SEVENTH EDITION

An Introduction to

Statistical Methods & Data Analysis

R. Lyman **Ott**
Michael **Longnecker**

An Introduction to

Statistical Methods & Data Analysis

An Introduction to Statistical Methods & Data Analysis

SEVENTH EDITION

R. Lyman Ott

Michael Longnecker

Texas A&M University

CENGAGE Learning®

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

CENGAGE Learning®

An Introduction to Statistical Methods and Data Analysis, Seventh Edition
R. Lyman Ott, Michael Longnecker

Senior Product Team Manager: Richard Stratton

Content Developer: Andrew Coppola

Associate Content Developer: Spencer Arritt

Product Assistant: Kathryn Schrumpf

Marketing Manager: Julie Schuster

Content Project Manager: Cheryll Linthicum

Art Director: Vernon Boes

Manufacturing Planner: Sandee Milewski

Intellectual Property Analyst: Christina Ciaramella

Intellectual Property Project Manager: Farah Fard

Production Service and Compositor: Cenveo Publishing Services

Photo and Text Researcher: Lumina Datamatics, LTD

Copy Editor:

Illustrator: Macmillan Publishing Services/ Cenveo Publishing Services

Text and Cover Designer: C. Miller

Cover Image: polygraphus/Getty Images

Library of Congress Control Number: 2015938496

ISBN: 978-1-305-26947-7

Cengage Learning
20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at **www.cengage.com**

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage Learning Solutions, visit **www.cengage.com**

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com**

Printed in the United States of America
Print Number: 01 Print Year: 2015

CONTENTS

PREFACE

INTENDED AUDIENCE

An Introduction to Statistical Methods and Data Analysis, Seventh Edition, provides a broad overview of statistical methods for advanced undergraduate and graduate students from a variety of disciplines. This book is intended to prepare students to solve problems encountered in research projects, to make decisions based on data in general settings both within and beyond the university setting, and finally to become critical readers of statistical analyses in research papers and in news reports. The book presumes that the students have a minimal mathematical background (high school algebra) and no prior course work in statistics. The first 11 chapters of the textbook present the material typically covered in an introductory statistics course. However, this book provides research studies and examples that connect the statistical concepts to data analysis problems that are often encountered in undergraduate capstone courses. The remaining chapters of the book cover regression modeling and design of experiments. We develop and illustrate the statistical techniques and thought processes needed to design a research study or experiment and then analyze the data collected using an intuitive and proven four-step approach. This should be especially helpful to graduate students conducting their MS thesis and PhD dissertation research.

MAJOR FEATURES OF TEXTBOOK

Learning from Data

In this text, we approach the study of statistics by considering a four-step process by which we can learn from data:

1. Defining the Problem
2. Collecting the Data
3. Summarizing the Data
4. Analyzing the Data, Interpreting the Analyses, and Communicating the Results

Case Studies

In order to demonstrate the relevance and critical nature of statistics in solving real-world problems, we introduce the major topic of each chapter using a case study. The case studies were selected from many sources to illustrate the broad applicability of statistical methodology. The four-step learning from data process is illustrated through the case studies. This approach will hopefully assist in overcoming

the natural initial perception held by many people that statistics is just another "math course." The introduction of major topics through the use of case studies provides a focus on the central nature of applied statistics in a wide variety of research and business-related studies. These case studies will hopefully provide the reader with an enthusiasm for the broad applicability of statistics and the statistical thought process that the authors have found and used through their many years of teaching, consulting, and R & D management. The following research studies illustrate the types of studies we have used throughout the text.

- **Exit Polls Versus Election Results:** A study of why the exit polls from 9 of 11 states in the 2004 presidential election predicted John Kerry as the winner when in fact President Bush won 6 of the 11 states.
- **Evaluation of the Consistency of Property Assessors:** A study to determine if county property assessors differ systematically in their determination of property values.
- **Effect of Timing of the Treatment of Port-Wine Stains with Lasers:** A prospective study that investigated whether treatment at a younger age would yield better results than treatment at an older age.
- **Controlling for Student Background in the Assessment of Teaching:** An examination of data used to support possible improvements to the No Child Left Behind program while maintaining the important concepts of performance standards and accountability.

Each of the research studies includes a discussion of the whys and hows of the study. We illustrate the use of the four-step learning from data process with each case study. A discussion of sample size determination, graphical displays of the data, and a summary of the necessary ingredients for a complete report of the statistical findings of the study are provided with many of the case studies.

Examples and Exercises

We have further enhanced the practical nature of statistics by using examples and exercises from journal articles, newspapers, and the authors' many consulting experiences. These will provide the students with further evidence of the practical usages of statistics in solving problems that are relevant to their everyday lives. Many new exercises and examples have been included in this edition of the book. The number and variety of exercises will be a great asset to both the instructor and students in their study of statistics.

Topics Covered

This book can be used for either a one-semester or a two-semester course. Chapters 1 through 11 would constitute a one-semester course. The topics covered would include

Chapter 1—Statistics and the scientific method
Chapter 2—Using surveys and experimental studies to gather data
Chapters 3 & 4—Summarizing data and probability distributions
Chapters 5–7—Analyzing data: inferences about central values and variances
Chapters 8 & 9—One-way analysis of variance and multiple comparisons

Chapter 10—Analyzing data involving proportions
Chapter 11—Linear regression and correlation

The second semester of a two-semester course would then include model building and inferences in multiple regression analysis, logistic regression, design of experiments, and analysis of variance:

Chapters 11–13—Regression methods and model building: multiple regression and the general linear model, logistic regression, and building regression models with diagnostics

Chapters 14–19—Design of experiments and analysis of variance: design concepts, analysis of variance for standard designs, analysis of covariance, random and mixed effects models, split-plot designs, repeated measures designs, crossover designs, and unbalanced designs

Emphasis on Interpretation, not Computation

In the book are examples and exercises that allow the student to study how to calculate the value of statistical estimators and test statistics using the definitional form of the procedure. After the student becomes comfortable with the aspects of the data the statistical procedure is reflecting, we then emphasize the use of computer software in making computations in the analysis of larger data sets. We provide output from three major statistical packages: SAS, Minitab, and SPSS. We find that this approach provides the student with the experience of computing the value of the procedure using the definition; hence, the student learns the basics behind each procedure. In most situations beyond the statistics course, the student should be using computer software in making the computations for both expedience and quality of calculation. In many exercises and examples, the use of the computer allows for more time to emphasize the interpretation of the results of the computations without having to expend enormous amounts of time and effort in the actual computations.

In numerous examples and exercises, the importance of the following aspects of hypothesis testing are demonstrated:

1. The statement of the research hypothesis through the summarization of the researcher's goals into a statement about population parameters.
2. The selection of the most appropriate test statistic, including sample size computations for many procedures.
3. The necessity of considering both Type I and Type II error rates (α and β) when discussing the results of a statistical test of hypotheses.
4. The importance of considering both the statistical significance and the practical significance of a test result. Thus, we illustrate the importance of estimating effect sizes and the construction of confidence intervals for population parameters.
5. The statement of the results of the statistical test in nonstatistical jargon that goes beyond the statement "reject H_0" or "fail to reject H_0."

New to the Seventh Edition

- There are instructions on the use of R code. R is a free software package that can be downloaded from **http:/ /lib.stat.cmu.edu/R/CRAN**.

Click your choice of platform (Linux, MacOS X, or Windows) for the precompiled binary distribution. Note the FAQs link to the left for additional information. Follow the instructions for installing the *base* system software (which is all you will need).

- New examples illustrate the breadth of applications of statistics to real-world problems.
- An alternative to the standard deviation, MAD, is provided as a measure of dispersion in a population/sample.
- The use of bootstrapping in obtaining confidence intervals and p-values is discussed.
- Instructions are included on how to use R code to obtain percentiles and probabilities from the following distributions: normal, binomial, Poisson, chi-squared, F, and t.
- A nonparametric alternative to the Pearson correlation coefficient: Spearman's rank correlation, is provided.
- The binomial test for small sample tests of proportions is presented.
- The McNemar test for paired count data has been added.
- The Akaike information criterion and Bayesian information criterion for variable selection are discussed.

Additional Features Retained from Previous Editions

- Many practical applications of statistical methods and data analysis from agriculture, business, economics, education, engineering, medicine, law, political science, psychology, environmental studies, and sociology have been included.
- The seventh edition contains over 1,000 exercises, with nearly 400 of the exercises new.
- Computer output from Minitab, SAS, and SPSS is provided in numerous examples. The use of computers greatly facilitates the use of more sophisticated graphical illustrations of statistical results.
- Attention is paid to the underlying assumptions. Graphical procedures and test procedures are provided to determine if assumptions have been violated. Furthermore, in many settings, we provide alternative procedures when the conditions are not met.
- The first chapter provides a discussion of "What Is Statistics?" We provide a discussion of why students should study statistics along with a discussion of several major studies that illustrate the use of statistics in the solution of real-life problems.

Ancillaries

- Student Solutions Manual (ISBN-10: 1-305-26948-9; ISBN-13: 978-1-305-26948-4), containing select worked solutions for problems in the textbook.
- A Companion Website at **www.cengage.com/statistics/ott,** containing downloadable data sets for Excel, Minitab, SAS, SPSS, and others, plus additional resources for students and faculty.

Acknowledgments

There are many people who have made valuable, constructive suggestions for the development of the original manuscript and during the preparation of the subsequent editions. We are very appreciative of the insightful and constructive comments from the following reviewers:

Naveen Bansal, Marquette University

Kameryn Denaro, San Diego State University

Mary Gray, American University

Craig Leth-Steensen, Carleton University

Jing Qian, University of Massachusetts

Mark Riggs, Abilene Christian University

Elaine Spiller, Marquette University

We are also appreciate of the preparation assistance received from Molly Taylor and Jay Campbell; the scheduling of the revisions by Mary Tindle, the Senior Project Manager at Cenveo Publisher Services, who made sure that the book was completed in a timely manner. The authors of the solutions manual, Soma Roy, California Polytechnic State University, and John Draper, The Ohio State University, provided me with excellent input which resulted in an improved set of exercises for the seventh edition. The person who assisted me the greatest degree in the preparation of the seventh edition, was Sherry Goldbecker, the copy editor. Sherry not only corrected my many grammatical errors but also provided rephrasing of many sentences which made for a more straight forward explanation of statistical concepts. The students, who use this book in their statistics classes, will be most appreciative of Sherry's many contributions.

PART 1

Introduction

CHAPTER 1 Statistics and the Scientific Method

CHAPTER 1

Statistics and the Scientific Method

1.1 Introduction

Statistics is the science of designing studies or experiments, collecting data, and modeling/analyzing data for the purpose of decision making and scientific discovery when the available information is both limited and variable. That is, statistics is the science of *Learning from Data.*

Almost everyone, including social scientists, medical researchers, superintendents of public schools, corporate executives, market researchers, engineers, government employees, and consumers, deals with data. These data could be in the form of quarterly sales figures, percent increase in juvenile crime, contamination levels in water samples, survival rates for patients undergoing medical therapy, census figures, or information that helps determine which brand of car to purchase. In this text, we approach the study of statistics by considering the four-step process in Learning from Data: (1) defining the problem, (2) collecting the data, (3) summarizing the data, and (4) analyzing the data, interpreting the analyses, and communicating the results. Through the use of these four steps in Learning from Data, our study of statistics closely parallels the Scientific Method, which is a set of principles and procedures used by successful scientists in their pursuit of knowledge. The method involves the formulation of research goals, the design of observational studies and/or experiments, the collection of data, the modeling/analysis of the data in the context of research goals, and the testing of hypotheses. The conclusion of these steps is often the formulation of new research goals for another study. These steps are illustrated in the schematic given in Figure 1.1.

This book is divided into sections corresponding to the four-step process in Learning from Data. The relationship among these steps and the chapters of the book is shown in Table 1.1. As you can see from this table, much time is spent discussing how to analyze data using the basic methods presented in Chapters 5–19.

FIGURE 1.1
Scientific Method Schematic

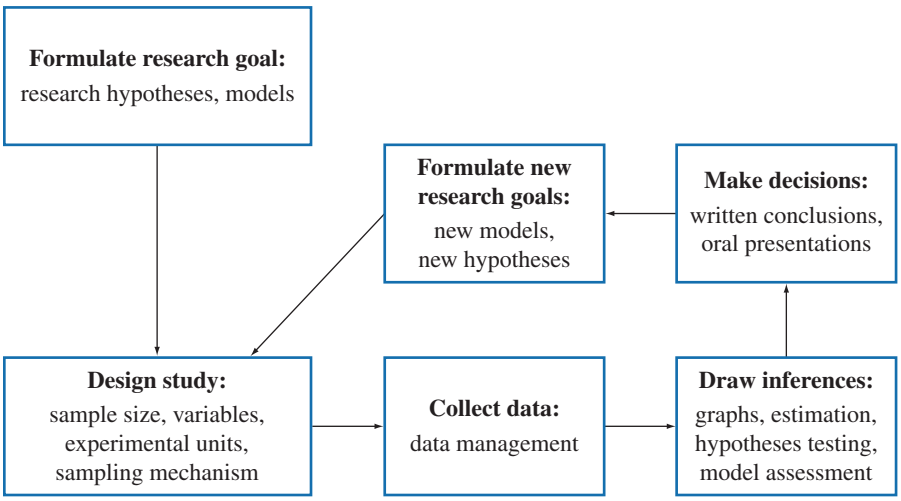


TABLE 1.1
Organization of the text

The Four-Step Process	Chapters
1 Defining the Problem	1 Statistics and the Scientific Method
2 Collecting the Data	2 Using Surveys and Experimental Studies to Gather Data
3 Summarizing the Data	3 Data Description
	4 Probability and Probability Distributions
4 Analyzing the Data, Interpreting the Analyses, and Communicating the Results	5 Inferences about Population Central Values
	6 Inferences Comparing Two Population Central Values
	7 Inferences about Population Variances
	8 Inferences about More Than Two Population Central Values
	9 Multiple Comparisons
	10 Categorical Data
	11 Linear Regression and Correlation
	12 Multiple Regression and the General Linear Model
	13 Further Regression Topics
	14 Analysis of Variance for Completely Randomized Designs
	15 Analysis of Variance for Blocked Designs
	16 The Analysis of Covariance
	17 Analysis of Variance for Some Fixed-, Random-, and Mixed-Effects Models
	18 Split-Plot, Repeated Measures, and Crossover Designs
	19 Analysis of Variance for Some Unbalanced Designs

However, you must remember that for each data set requiring analysis, someone has defined the problem to be examined (Step 1), developed a plan for collecting data to address the problem (Step 2), and summarized the data and prepared the data for analysis (Step 3). Then following the analysis of the data, the results of the analysis must be interpreted and communicated either verbally or in written form to the intended audience (Step 4).

All four steps are important in Learning from Data; in fact, unless the problem to be addressed is clearly defined and the data collection carried out properly, the interpretation of the results of the analyses may convey misleading information because the analyses were based on a data set that did not address the problem or that was incomplete and contained improper information. Throughout the text,

we will try to keep you focused on the bigger picture of Learning from Data through the four-step process. Most chapters will end with a summary section that emphasizes how the material of the chapter fits into the study of statistics—Learning from Data.

To illustrate some of the above concepts, we will consider four situations in which the four steps in Learning from Data could assist in solving a real-world problem.

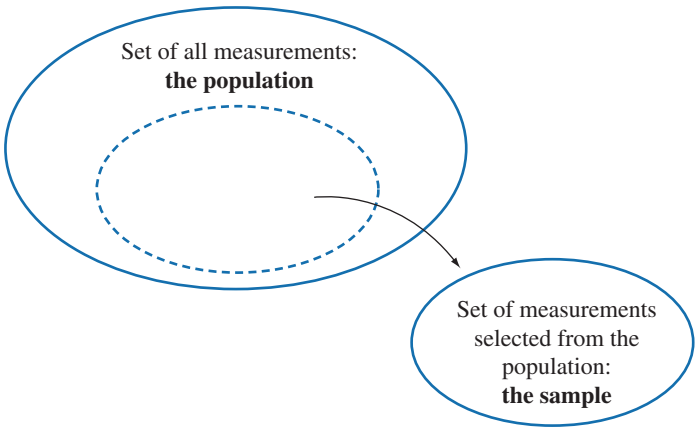
1. **Problem: Inspection of ground beef in a large beef-processing facility.** A beef-processing plant produces approximately half a million packages of ground beef per week. The government inspects packages for possible improper labeling of the packages with respect to the percent fat in the meat. The inspectors must open the ground beef package in order to determine the fat content of the ground beef. The inspection of every package would be prohibitively costly and time consuming. An alternative approach is to select 250 packages for inspection from the daily production of 100,000 packages. The fraction of packages with improper labeling in the sample of 250 packages would then be used to estimate the fraction of packages improperly labeled in the complete day's production. If this fraction exceeds a set specification, action is then taken against the meat processor. In later chapters, a procedure will be formulated to determine how well the sample fraction of improperly labeled packages approximates the fraction of improperly labeled packages for the whole day's output.
2. **Problem: Is there a relationship between quitting smoking and gaining weight?** To investigate the claim that people who quit smoking often experience a subsequent weight gain, researchers selected a random sample of 400 participants who had successfully participated in programs to quit smoking. The individuals were weighed at the beginning of the program and again 1 year later. The average change in weight of the participants was an increase of 5 pounds. The investigators concluded that there was evidence that the claim was valid. We will develop techniques in later chapters to assess when changes are truly significant changes and not changes due to random chance.
3. **Problem: What effect does nitrogen fertilizer have on wheat production?** For a study of the effects of nitrogen fertilizer on wheat production, a total of 15 fields was available to the researcher. She randomly assigned three fields to each of the five nitrogen rates under investigation. The same variety of wheat was planted in all 15 fields. The fields were cultivated in the same manner until harvest, and the number of pounds of wheat per acre was then recorded for each of the 15 fields. The experimenter wanted to determine the optimal level of nitrogen to apply to *any* wheat field, but, of course, she was limited to running experiments on a limited number of fields. After determining the amount of nitrogen that yielded the largest production of wheat in the study fields, the experimenter then concluded that similar results would hold for wheat fields possessing characteristics somewhat the same as the study fields. Is the experimenter justified in reaching this conclusion?

4. **Problem: Determining public opinion toward a question, issue, product, or candidate.** Similar applications of statistics are brought to mind by the frequent use of the *New York Times/CBS News, Washington Post/ABC News, Wall Street Journal/NBC News, Harris, Gallup/Newsweek,* and *CNN/Time* polls. How can these pollsters determine the opinions of more than 195 million Americans who are of voting age? They certainly do not contact every potential voter in the United States. Rather, they sample the opinions of a small number of potential voters, perhaps as few as 1,500, to estimate the reaction of every person of voting age in the country. The amazing result of this process is that if the selection of the voters is done in an unbiased way and voters are asked unambiguous, nonleading questions, the fraction of those persons contacted who hold a particular opinion will closely match the fraction in the total population holding that opinion at a particular time. We will supply convincing supportive evidence of this assertion in subsequent chapters.

These problems illustrate the four-step process in Learning from Data. First, there was a problem or question to be addressed. Next, for each problem a study or experiment was proposed to collect meaningful data to solve the problem. The government meat inspection agency had to decide both how many packages to inspect per day and how to select the sample of packages from the total daily output in order to obtain a valid prediction. The polling groups had to decide how many voters to sample and how to select these individuals in order to obtain information that is representative of the population of all voters. Similarly, it was necessary to carefully plan how many participants in the weight-gain study were needed and how they were to be selected from the list of all such participants. Furthermore, what variables did the researchers have to measure on each participant? Was it necessary to know each participant's age, sex, physical fitness, and other health-related variables, or was weight the only important variable? The results of the study may not be relevant to the general population if many of the participants in the study had a particular health condition. In the wheat experiment, it was important to measure both the soil characteristics of the fields and the environmental conditions, such as temperature and rainfall, to obtain results that could be generalized to fields not included in the study. The design of a study or experiment is crucial to obtaining results that can be generalized beyond the study.

Finally, having collected, summarized, and analyzed the data, it is important to report the results in unambiguous terms to interested people. For the meat inspection example, the government inspection agency and the personnel in the beef-processing plant would need to know the distribution of fat content in the daily production of ground beef. Based on this distribution, the agency could then impose fines or take other remedial actions against the production facility. Also, knowledge of this distribution would enable company production personnel to make adjustments to the process in order to obtain acceptable fat content in their ground beef packages. Therefore, the results of the statistical analyses cannot be presented in ambiguous terms; decisions must be made from a well-defined knowledge base. The results of the weight-gain study would be of vital interest to physicians who have patients participating in the smoking-cessation program. If a significant increase in weight was recorded for those individuals who had quit smoking, physicians would have to recommend diets so that the former smokers

FIGURE 1.2
Population and sample



would not go from one health problem (smoking) to another (elevated blood pressure due to being overweight). It is crucial that a careful description of the participants—that is, age, sex, and other health-related information—be included in the report. In the wheat study, the experiment would provide farmers with information that would allow them to economically select the optimum amount of nitrogen required for their fields. Therefore, the report must contain information concerning the amount of moisture and types of soils present on the study fields. Otherwise, the conclusions about optimal wheat production may not pertain to farmers growing wheat under considerably different conditions.

population

sample

To infer validly that the results of a study are applicable to a larger group than just the participants in the study, we must carefully define the **population** (see Definition 1.1) to which inferences are sought and design a study in which the **sample** (see Definition 1.2) has been appropriately selected from the designated population. We will discuss these issues in Chapter 2.

DEFINITION 1.1

A **population** is the set of all measurements of interest to the sample collector. (See Figure 1.2.)

DEFINITION 1.2

A **sample** is any subset of measurements selected from the population. (See Figure 1.2.)

1.2 Why Study Statistics?

We can think of many reasons for taking an introductory course in statistics. One reason is that you need to know how to evaluate published numerical facts. Every person is exposed to manufacturers' claims for products; to the results of sociological, consumer, and political polls; and to the published results of scientific research. Many of these results are inferences based on sampling. Some inferences are valid; others are invalid. Some are based on samples of adequate size; others are not. Yet all these published results bear the ring of truth. Some people (particularly statisticians) say that statistics can be made to support almost

anything. Others say it is easy to lie with statistics. Both statements are true. It is easy, purposely or unwittingly, to distort the truth by using statistics when presenting the results of sampling to the uninformed. It is thus crucial that you become an informed and critical reader of data-based reports and articles.

A second reason for studying statistics is that your profession or employment may require you to interpret the results of sampling (surveys or experimentation) or to employ statistical methods of analysis to make inferences in your work. For example, practicing physicians receive large amounts of advertising describing the benefits of new drugs. These advertisements frequently display the numerical results of experiments that compare a new drug with an older one. Do such data really imply that the new drug is more effective, or is the observed difference in results due simply to random variation in the experimental measurements?

Recent trends in the conduct of court trials indicate an increasing use of probability and statistical inference in evaluating the quality of evidence. The use of statistics in the social, biological, and physical sciences is essential because all these sciences make use of observations of natural phenomena, through sample surveys or experimentation, to develop and test new theories. Statistical methods are employed in business when sample data are used to forecast sales and profit. In addition, they are used in engineering and manufacturing to monitor product quality. The sampling of accounts is a useful tool to assist accountants in conducting audits. Thus, statistics plays an important role in almost all areas of science, business, and industry; persons employed in these areas need to know the basic concepts, strengths, and limitations of statistics.

The article ***"What Educated Citizens Should Know About Statistics and Probability," by J. Utts (2003),*** contains a number of statistical ideas that need to be understood by users of statistical methodology in order to avoid confusion in the use of their research findings. Misunderstandings of statistical results can lead to major errors by government policymakers, medical workers, and consumers of this information. The article selected a number of topics for discussion. We will summarize some of the findings in the article. A complete discussion of all these topics will be given throughout the book.

1. One of the most frequent misinterpretations of statistical findings is when a statistically significant relationship is established between two variables and it is then concluded that a change in the explanatory variable *causes* a change in the response variable. As will be discussed in the book, this conclusion can be reached only under very restrictive constraints on the experimental setting. Utts examined a recent *Newsweek* article discussing the relationship between the strength of religious beliefs and physical healing. Utts' article discussed the problems in reaching the conclusion that the stronger a patient's religious beliefs, the more likely the patient would be cured of his or her ailment. Utts showed that there are numerous other factors involved in a patient's health and the conclusion that religious beliefs cause a cure cannot be validly reached.
2. A common confusion in many studies is the difference between *(statistically) significant* findings in a study and *(practically) significant* findings. This problem often occurs when large data sets are involved in a study or experiment. This type of problem will be discussed in detail throughout the book. We will use a number of examples that will illustrate how this type of confusion can be avoided by researchers when reporting the findings of their experimental results.

Utts' article illustrated this problem with a discussion of a study that found a statistically significant difference in the average heights of military recruits born in the spring and in the fall. There were 507,125 recruits in the study and the difference in average height was about 1/4 inch. So, even though there may be a difference in the actual average heights of recruits in the spring and the fall, the difference is so small (1/4 inch) that it is of no practical importance.

3. The size of the sample also may be a determining factor in studies in which statistical significance is *not* found. A study may not have selected a sample size large enough to discover a difference between the several populations under study. In many government-sponsored studies, the researchers do not receive funding unless they are able to demonstrate that the sample sizes selected for their study are of an appropriate size to detect specified differences in populations if in fact they exist. Methods to determine appropriate sample sizes will be provided in the chapters on hypotheses testing and experimental design.
4. Surveys are ubiquitous, especially during the years in which national elections are held. In fact, market surveys are nearly as widespread as political polls. There are many sources of bias that can creep into the most reliable of surveys. The manner in which people are selected for inclusion in the survey, the way in which questions are phrased, and even the manner in which questions are posed to the subject may affect the conclusions obtained from the survey. We will discuss these issues in Chapter 2.
5. Many students find the topic of probability to be very confusing. One of these confusions involves conditional probability where the probability of an event occurring is computed under the condition that a second event has occurred with certainty. For example, a new diagnostic test for the pathogen *Escherichia coli* in meat is proposed to the U.S. Department of Agriculture (USDA). The USDA evaluates the test and determines that the test has both a low *false positive* rate and a low *false negative* rate. That is, it is very unlikely that the test will declare the meat contains *E. coli* when in fact it does not contain *E. coli.* Also, it is very unlikely that the test will declare the meat does not contain *E. coli* when in fact it does contain *E. coli.* Although the diagnostic test has a very low false positive rate and a very low false negative rate, the probability that *E. coli* is in fact present in the meat when the test yields a positive test result is *very* low for those situations in which a particular strain of *E. coli* occurs very infrequently. In Chapter 4, we will demonstrate how this probability can be computed in order to provide a true assessment of the performance of a diagnostic test.
6. Another concept that is often misunderstood is the role of the degree of variability in interpreting what is a "normal" occurrence of some naturally occurring event. Utts' article provided the following example. A company was having an odor problem with its wastewater treatment plant. It attributed the problem to "abnormal" rainfall during the period in which the odor problem was occurring. A company official stated that the facility experienced 170% to 180% of its "normal" rainfall during this period, which resulted in the water in

the holding ponds taking longer to exit for irrigation. Thus, there was more time for the pond to develop an odor. The company official did not point out that yearly rainfall in this region is extremely variable. In fact, the historical range for rainfall is between 6.1 and 37.4 inches with a median rainfall of 16.7 inches. The rainfall for the year of the odor problem was 29.7 inches, which was well within the "normal" range for rainfall. There was a confusion between the terms "average" and "normal" rainfall. The concept of natural variability is crucial to correct interpretation of statistical results. In this example, the company official should have evaluated the percentile for an annual rainfall of 29.7 inches in order to demonstrate the abnormality of such a rainfall. We will discuss the ideas of data summaries and percentiles in Chapter 3.

The types of problems expressed above and in Utts' article represent common and important misunderstandings that can occur when researchers use statistics in interpreting the results of their studies. We will attempt throughout the book to discuss possible misinterpretations of statistical results and how to avoid them in your data analyses. More importantly, we want the reader of this book to become a discriminating reader of statistical findings, the results of surveys, and project reports.

1.3 Some Current Applications of Statistics

Defining the Problem: Obtaining Information from Massive Data Sets

Data mining is defined to be a process by which useful information is obtained from large sets of data. Data mining uses statistical techniques to discover patterns and trends that are present in a large data set. In most data sets, important patterns would not be discovered by using traditional data exploration techniques because the types of relationships between the many variables in the data set are either too complex or because the data sets are so large that they mask the relationships.

The patterns and trends discovered in the analysis of the data are defined as data mining models. These models can be applied to many different situations, such as:

- Forecasting: Estimating future sales, predicting demands on a power grid, or estimating server downtime
- Assessing risk: Choosing the rates for insurance premiums, selecting best customers for a new sales campaign, determining which medical therapy is most appropriate given the physiological characteristics of the patient
- Identifying sequences: Determining customer preferences in online purchases, predicting weather events
- Grouping: Placing customers or events into cluster of related items, analyzing and predicting relationships between demographic characteristics and purchasing patterns, identifying fraud in credit card purchases

A new medical procedure referred to as gene editing has the potential to assist thousands of people suffering many different diseases. An article in the ***Houston Chronicle (2013)***, describes how data mining techniques are used to

explore massive genomic data bases to interpret millions of bits of data in a person's DNA. This information is then used to identify a single defective gene, which is cut out, and splice in a correction. This area of research is referred to as biomedical informatics and is based on the premise that the human body is a data bank of incredible depth and complexity. It is predicted that by 2015, the average hospital will have approximately 450 terabytes of patient data consisting of large, complex images from CT scans, MRIs, and other imaging techniques. However, only a small fraction of the current medical data has been analyzed, thus opening huge opportunities for persons trained in data mining. In a case described in the article, a 7-year-old boy tormented by scabs, blisters, and scars was given a new lease on life by using data mining techniques to discover a single letter in his faulty genome.

Defining the Problem: Determining the Effectiveness of a New Drug Product

The development and testing of the Salk vaccine for protection against poliomyelitis (polio) provide an excellent example of how statistics can be used in solving practical problems. Most parents and children growing up before 1954 can recall the panic brought on by the outbreak of polio cases during the summer months. Although relatively few children fell victim to the disease each year, the pattern of outbreak of polio was unpredictable and caused great concern because of the possibility of paralysis or death. The fact that very few of today's youth have even heard of polio demonstrates the great success of the vaccine and the testing program that preceded its release on the market.

It is standard practice in establishing the effectiveness of a particular drug product to conduct an experiment (often called a *clinical trial*) with human participants. For some clinical trials, assignments of participants are made at random, with half receiving the drug product and the other half receiving a solution or tablet that does not contain the medication (called a *placebo*). One statistical problem concerns the determination of the total number of participants to be included in the clinical trial. This problem was particularly important in the testing of the Salk vaccine because data from previous years suggested that the incidence rate for polio might be less than 50 cases for every 100,000 children. Hence, a large number of participants had to be included in the clinical trial in order to detect a difference in the incidence rates for those treated with the vaccine and those receiving the placebo.

With the assistance of statisticians, it was decided that a total of 400,000 children should be included in the Salk clinical trial begun in 1954, with half of them randomly assigned the vaccine and the remaining children assigned the placebo. No other clinical trial had ever been attempted on such a large group of participants. Through a public school inoculation program, the 400,000 participants were treated and then observed over the summer to determine the number of children contracting polio. Although fewer than 200 cases of polio were reported for the 400,000 participants in the clinical trial, more than three times as many cases appeared in the group receiving the placebo. These results, together with some statistical calculations, were sufficient to indicate the effectiveness of the Salk polio vaccine. However, these conclusions would not have been possible if the statisticians and scientists had not planned for and conducted such a large clinical trial.

The development of the Salk vaccine is not an isolated example of the use of statistics in the testing and development of drug products. In recent years,

the U.S. Food and Drug Administration (FDA) has placed stringent requirements on pharmaceutical firms wanting to establish the effectiveness of proposed new drug products. Thus, statistics has played an important role in the development and testing of birth control pills, rubella vaccines, chemotherapeutic agents in the treatment of cancer, and many other preparations.

Defining the Problem: Improving the Reliability of Evidence in Criminal Investigations

The ***National Academy of Sciences*** released a report (National Research Council, 2009) in which one of the more important findings was the need for applying statistical methods in the design of studies used to evaluate inferences from evidence gathered by forensic technicians. The following statement is central to the report:

> "Over the last two decades, advances in some forensic science disciplines, especially the use of DNA technology, have demonstrated that some areas of forensic science have great additional potential to help law enforcement identify criminals.... Those advances, however, also have revealed that, in some cases, substantive information and testimony based on faulty forensic science analyses may have contributed to wrongful convictions of innocent people. This fact has demonstrated the potential danger of giving undue weight to evidence and testimony derived from imperfect testing and analysis."

There are many sources that may impact the accuracy of conclusions inferred from the crime scene evidence and presented to a jury by a forensic investigator. Statistics can play a role in improving forensic analyses. Statistical principles can be used to identify sources of variation and quantify the size of the impact that these sources of variation can have on the conclusions reached by the forensic investigator.

An illustration of the impact of an inappropriately designed study and statistical analysis on the conclusions reached from the evidence obtained at a crime scene can be found in ***Spiegelman et al. (2007)***. They demonstrate that the evidence used by the FBI crime lab to support the claim that there was not a second assassin of President John F. Kennedy was based on a faulty analysis of the data and an overstatement of the results of a method of forensic testing called Comparative Bullet Lead Analysis (CBLA). This method applies a chemical analysis to link a bullet found at a crime scene to the gun that had discharged the bullet. Based on evidence from chemical analyses of the recovered bullet fragments, the 1979 U.S. House Select Committee on Assassinations concluded that all the bullets striking President Kennedy were fired from Lee Oswald's rifle. A new analysis of the bullets using more appropriate statistical analyses demonstrated that the evidence presented in 1979 was overstated. A case is presented for a new analysis of the assassination bullet fragments, which may shed light on whether the five bullet fragments found in the Kennedy assassination are derived from three or more bullets and not just two bullets, as was presented as the definitive evidence that Oswald was the sole shooter in the assassination of President Kennedy.

Defining the Problem: Estimating Bowhead Whale Population Size

Raftery and Zeh (1998) discuss the estimation of the population size and rate of increase in bowhead whales, *Balaena mysticetus.* The importance of such a study derives from the fact that bowheads were the first species of great whale for

which commercial whaling was stopped; thus, their status indicates the recovery prospects of other great whales. Also, the International Whaling Commission uses these estimates to determine the aboriginal subsistence whaling quota for Alaskan Eskimos. To obtain the necessary data, researchers conducted a visual and acoustic census off Point Barrow, Alaska. The researchers then applied statistical models and estimation techniques to the data obtained in the census to determine whether the bowhead population had increased or decreased since commercial whaling was stopped. The statistical estimates showed that the bowhead population was increasing at a healthy rate, indicating that stocks of great whales that have been decimated by commercial hunting can recover after hunting is discontinued.

Defining the Problem: Ozone Exposure and Population Density

Ambient ozone pollution in urban areas is one of the nation's most pervasive environmental problems. Whereas the decreasing stratospheric ozone layer may lead to increased instances of skin cancer, high ambient ozone intensity has been shown to cause damage to the human respiratory system as well as to agricultural crops and trees. The Houston, Texas, area has ozone concentrations and are rated second only to those of Los Angeles. that exceed the National Ambient Air Quality Standard. ***Carroll et al. (1997)*** describe how to analyze the hourly ozone measurements collected in Houston from 1980 to 1993 by 9 to 12 monitoring stations. Besides the ozone level, each station recorded three meteorological variables: temperature, wind speed, and wind direction.

The statistical aspect of the project had three major goals:

1. Provide information (and/or tools to obtain such information) about the amount and pattern of missing data as well as about the quality of the ozone and the meteorological measurements.
2. Build a model of ozone intensity to predict the ozone concentration at any given location within Houston at any given time between 1980 and 1993.
3. Apply this model to estimate exposure indices that account for either a long-term exposure or a short-term high-concentration exposure; also, relate census information to different exposure indices to achieve population exposure indices.

The spatial–temporal model the researchers built provided estimates demonstrating that the highest ozone levels occurred at locations with relatively small populations of young children. Also, the model estimated that the exposure of young children to ozone decreased by approximately 20% from 1980 to 1993. An examination of the distribution of population exposure had several policy implications. In particular, it was concluded that the current placement of monitors is not ideal if one is concerned with assessing population exposure. This project involved all four components of Learning from Data: planning where the monitoring stations should be placed within the city, how often the data should be collected, and what variables should be recorded; conducting spatial–temporal graphing of the data; creating spatial–temporal models of the ozone data, meteorological data, and demographic data; and, finally, writing a report that could assist local and federal officials in formulating policy with respect to decreasing ozone levels.

Defining the Problem: Assessing Public Opinion

Public opinion, consumer preference, and election polls are commonly used to assess the opinions or preferences of a segment of the public regarding issues, products, or candidates of interest. We, the American public, are exposed to the results of these polls daily in newspapers, in magazines, on the internet, on the radio, and on television. For example, the results of polls related to the following subjects were printed in local newspapers:

- Public confidence in the potential for job growth in the coming year
- Reactions of Texas residents to the state legislature's failure to expand Medicaid coverage
- Voters' preferences for tea party candidates in the fall congressional elections
- Attitudes toward increasing the gasoline tax in order to increase funding for road construction and maintenance
- Product preference polls related to specific products (Toyota vs. Ford, DirecTV vs. Comcast, Dell vs. Apple, Subway vs. McDonald's)
- Public opinion on a national immigration policy

A number of questions can be raised about polls. Suppose we consider a poll on the public's opinion on a proposed income tax increase in the state of Michigan. *What was the population of interest to the pollster?* Was the pollster interested in all residents of Michigan or just those citizens who currently pay income taxes? *Was the sample in fact selected from this population?* If the population of interest was all persons currently paying income taxes, did the pollster make sure that all the individuals sampled were current taxpayers? *What questions were asked and how were the questions phrased?* Was each person asked the same question? Were the questions phrased in such a manner as to bias the responses? Can we believe the results of these polls? Do these results "represent" how the general public *currently* feels about the issues raised in the polls?

Opinion and preference polls are an important, visible application of statistics for the consumer. We will discuss this topic in more detail in Chapters 2 and 10. We hope that after studying this material you will have a better understanding of how to interpret the results of these polls.

1.4 A Note to the Student

We think with words and concepts. A study of the discipline of statistics requires us to memorize new terms and concepts (as does the study of a foreign language). Commit these definitions, theorems, and concepts to memory.

Also, focus on the broader concept of making sense of data. Do not let details obscure these broader characteristics of the subject. The teaching objective of this text is to identify and amplify these broader concepts of statistics.

1.5 Summary

The discipline of statistics and those who apply the tools of that discipline deal with Learning from Data. Medical researchers, social scientists, accountants, agronomists, consumers, government leaders, and professional statisticians are all involved with data collection, data summarization, data analysis, and the effective communication of the results of data analysis.

1.6 Exercises

1.1 Introduction

Bio. **1.1** ***Hansen (2006)*** describes a study to assess the migration and survival of salmon released from fish farms located in Norway. The mingling of escaped farmed salmon with wild salmon raises several concerns. First, the assessment of the abundance of wild salmon stocks will be biased if there is a presence of large numbers of farmed salmon. Second, potential interbreeding between farmed and wild salmon may result in a reduction in the health of the wild stocks. Third, diseases present in farmed salmon may be transferred to wild salmon. Two batches of farmed salmon were tagged and released in two locations, one batch of 1,996 fish in northern Norway and a second batch of 2,499 fish in southern Norway. The researchers recorded the time and location at which the fish were captured by either commercial fisherman or anglers in fresh water. Two of the most important pieces of information to be determined by the study were the distance from the point of the fish's release to the point of its capture and the length of time it took for the fish to be captured.

- **a.** Identify the population that is of interest to the researchers.
- **b.** Describe the sample.
- **c.** What characteristics of the population are of interest to the researchers?
- **d.** If the sample measurements are used to make inferences about the population characteristics, why is a measure of reliability of the inferences important?

Env. **1.2** During 2012, Texas had listed on FracFocus, an industry fracking disclosure site, nearly 6,000 oil and gas wells in which the fracking methodology was used to extract natural gas. ***Fontenot et al. (2013)*** reports on a study of 100 private water wells in or near the Barnett Shale in Texas. There were 91 private wells located within 5 km of an active gas well using fracking, 4 private wells with no gas wells located within a 14 km radius, and 5 wells outside of the Barnett Shale with no gas well located with a 60 km radius. They found that there were elevated levels of potential contaminants such as arsenic and selenium in the 91 wells closest to natural gas extraction sites compared to the 9 wells that were at least 14 km away from an active gas well using the £racking technique to extract natural gas.

- **a.** Identify the population that is of interest to the researchers.
- **b.** Describe the sample.
- **c.** What characteristics of the population are of interest to the researchers?
- **d.** If the sample measurements are used to make inferences about the population characteristics, why is a measure of reliability of the inferences important?

Soc. **1.3** In 2014, Congress cut $8.7 billion from the Supplemental Nutrition Assistance Program (SNAP), more commonly referred to as food stamps. The rationale for the decrease is that providing assistance to people will result in the next generation of citizens being more dependent on the government for support. ***Hoynes (2012)*** describes a study to evaluate this claim. The study examines 60,782 families over the time period of 1968 to 2009 which is subsequent to the introduction of the Food Stamp Program in 1961. This study examines the impact of a positive and policy-driven change in economic resources available in utero and during childhood on the economic health of individuals in adulthood. The study assembled data linking family background in early childhood to adult health and economic outcomes. The study concluded that the Food Stamp Program has effects decades after initial exposure. Specifically, access to food stamps in childhood leads to a significant reduction in the incidence of metabolic syndrome (obesity, high blood pressure, and diabetes) and, for women, an increase in economic self-sufficiency. Overall, the results suggest substantial internal and external benefits of SNAP.

- **a.** Identify the population that is of interest to the researchers.
- **b.** Describe the sample.
- **c.** What characteristics of the population are of interest to the researchers?
- **d.** If the sample measurements are used to make inferences about the population characteristics, why is a measure of reliability of the inferences important?

Med. **1.4** Of all sports, football accounts for the highest incidence of concussion in the United States due to the large number of athletes participating and the nature of the sport. While there is general agreement that concussion incidence can be reduced by making rule changes and teaching proper tackling technique, there remains debate as to whether helmet design may also reduce the incidence of concussion. ***Rowson et al. (2014)*** report on a retrospective analysis of head impact data collected between 2005 and 2010 from eight collegiate football teams. Concussion rates for players wearing two types of helmets, Riddell VSR4 and Riddell Revolution, were compared. A total of 1,281,444 head impacts were recorded, from which 64 concussions were diagnosed. The relative risk of sustaining a concussion in a Revolution helmet compared with a VSR4 helmet was 46.1%. This study illustrates that differences in the ability to reduce concussion risk exist between helmet models in football. Although helmet design may never prevent all concussions from occurring in football, evidence illustrates that it can reduce the incidence of this injury.

- **a.** Identify the population that is of interest to the researchers.
- **b.** Describe the sample.
- **c.** What characteristics of the population are of interest to the researchers?
- **d.** If the sample measurements are used to make inferences about the population characteristics, why is a measure of reliability of the inferences important?

Pol. Sci. **1.5** During the 2004 senatorial campaign in a large southwestern state, illegal immigration was a major issue. One of the candidates argued that illegal immigrants made use of educational and social services without having to pay property taxes. The other candidate pointed out that the cost of new homes in their state was 20–30% less than the national average due to the low wages received by the large number of illegal immigrants working on new home construction. A random sample of 5,500 registered voters was asked the question, "Are illegal immigrants generally a benefit or a liability to the state's economy?" The results were as follows: 3,500 people responded "liability," 1,500 people responded "benefit," and 500 people responded "uncertain."

- **a.** What is the population of interest?
- **b.** What is the population from which the sample was selected?
- **c.** Does the sample adequately represent the population?
- **d.** If a second random sample of 5,000 registered voters was selected, would the results be nearly the same as the results obtained from the initial sample of 5,000 voters? Explain your answer.

Edu. **1.6** An American history professor at a major university was interested in knowing the history literacy of college freshmen. In particular, he wanted to find what proportion of college freshmen at the university knew which country controlled the original 13 colonies prior to the American Revolution. The professor sent a questionnaire to all freshman students enrolled in HIST 101 and received responses from 318 students out of the 7,500 students who were sent the questionnaire. One of the questions was "What country controlled the original 13 colonies prior to the American Revolution?"

- **a.** What is the population of interest to the professor?
- **b.** What is the sampled population?
- **c.** Is there a major difference in the two populations. Explain your answer.
- **d.** Suppose that several lectures on the American Revolution had been given in HIST 101 prior to the students receiving the questionnaire. What possible source of bias has the professor introduced into the study relative to the population of interest?

PART 2

Collecting Data

CHAPTER 2 Using Surveys and Experimental Studies to Gather Data

CHAPTER 2

Using Surveys and Experimental Studies to Gather Data

2.1 Introduction and Abstract of Research Study

As mentioned in Chapter 1, the first step in Learning from Data is to define the problem. The design of the data collection process is the crucial step in *intelligent data gathering.* The process takes a conscious, concerted effort focused on the following steps:

- Specifying the objective of the study, survey, or experiment
- Identifying the variable(s) of interest
- Choosing an appropriate design for the survey or experimental study
- Collecting the data

To specify the objective of the study, you must understand the problem being addressed. For example, the transportation department in a large city wants to assess the public's perception of the city's bus system in order to increase the use of buses within the city. Thus, the department needs to determine what aspects of the bus system determine whether or not a person will ride the bus. The objective of the study is to identify factors that the transportation department can alter to increase the number of people using the bus system.

To identify the variables of interest, you must examine the objective of the study. For the bus system, some major factors can be identified by reviewing studies conducted in other cities and by brainstorming with the bus system employees. Some of the factors may be safety, cost, cleanliness of the buses, whether or not there is a bus stop close to the person's home or place of employment, and how often the bus fails to be on time. The measurements to be obtained in the

study would consist of importance ratings (very important, important, no opinion, somewhat unimportant, very unimportant) of the identified factors. Demographic information, such as age, sex, income, and place of residence, would also be measured. Finally, the measurement of variables related to how frequently a person currently rides the buses would be of importance. Once the objectives are determined and the variables of interest are specified, you must select the most appropriate method to collect the data. Data collection processes include surveys, experiments, and the examination of existing data from business records, censuses, government records, and previous studies. The theory of sample surveys and the theory of experimental designs provide excellent methodology for data collection. Usually surveys are passive. The goal of the survey is to gather data on existing conditions, attitudes, or behaviors. Thus, the transportation department would need to construct a questionnaire and then sample current riders of the buses and persons who use other forms of transportation within the city.

Experimental studies, on the other hand, tend to be more active: The person conducting the study varies the experimental conditions to study the effect of the conditions on the outcome of the experiment. For example, the transportation department could decrease the bus fares on a few selected routes and assess whether the use of its buses increased. However, in this example, other factors not under the bus system's control may also have changed during this time period. Thus, an increase in bus use may have taken place because of a strike of subway workers or an increase in gasoline prices. The decrease in fares was only one of several factors that may have "caused" the increase in the number of persons riding the buses.

In most experimental studies, as many as possible of the factors that affect the measurements are under the control of the experimenter. A floriculturist wants to determine the effect of a new plant stimulator on the growth of a commercially produced flower. The floriculturist would run the experiments in a greenhouse, where temperature, humidity, moisture levels, and sunlight are controlled. An equal number of plants would be treated with each of the selected quantities of the growth stimulator, including a control—that is, no stimulator applied. At the conclusion of the experiment, the size and health of the plants would be measured. The optimal level of the plant stimulator could then be determined because ideally all other factors affecting the size and health of the plants would be the same for all plants in the experiment.

In this chapter, we will consider some sampling designs for surveys and some designs for experimental studies. We will also make a distinction between an experimental study and an observational study.

Abstract of Research Study: Exit Polls Versus Election Results

As the 2004 presidential campaign approached Election Day, the Democratic Party was very optimistic that its candidate, John Kerry, would defeat the incumbent, George Bush. Many Americans arrived home the evening of Election Day to watch or listen to the network coverage of the election with the expectation that John Kerry would be declared the winner of the presidential race because throughout Election Day, radio and television reporters had provided exit poll results showing John Kerry ahead in nearly every crucial state, and in many of these states leading by substantial margins. The Democratic Party, being better organized with a greater commitment and focus than in many previous presidential elections, had produced an enormous number of Democratic loyalists for this election. But, as

TABLE 2.1 Predicted vs. actual percentages in battleground states

Crucial State	Sample	Exit Poll Results			Election Results			Election vs. Exit
		Bush	Kerry	Difference	Bush	Kerry	Difference	
Colorado	2,515	49.9%	48.1%	Bush 1.8%	52.0%	46.8%	Bush 5.2%	Bush 3.4%
Florida	2,223	48.8%	49.2%	Kerry 0.4%	49.4%	49.8%	Kerry 0.4%	No Diff.
Iowa	2,846	49.8%	49.7%	Bush 0.1%	52.1%	47.1%	Bush 5.0%	Bush 4.9%
Michigan	2,502	48.4%	49.7%	Kerry 1.3%	50.1%	49.2%	Bush 0.9%	Bush 2.2%
Minnesota	2,452	46.5%	51.1%	Kerry 4.6%	47.8%	51.2%	Kerry 3.4%	Kerry 1.2%
Nevada	2,178	44.5%	53.5%	Kerry 9.0%	47.6%	51.1%	Kerry 3.5%	Kerry 5.5%
New Hampshire	2,116	47.9%	49.2%	Kerry 1.3%	50.5%	47.9%	Bush 2.6%	Bush 3.9%
New Mexico	1,849	44.1%	54.9%	Kerry 10.8%	49.0%	50.3%	Kerry 1.3%	Kerry 9.5%
Ohio	1,951	47.5%	50.1%	Kerry 2.6%	50.0%	48.9%	Bush 1.1%	Bush 3.7%
Pennsylvania	1,963	47.9%	52.1%	Kerry 4.2%	51.0%	48.5%	Bush 2.5%	Bush 6.7%
Wisconsin	1,930	45.4%	54.1%	Kerry 8.7%	48.6%	50.8%	Kerry 2.2%	Kerry 6.5%

the evening wore on, in one crucial state after another the election returns showed results that differed greatly from what the exit polls had predicted.

The data shown in Table 2.1 are from a University of Pennsylvania technical report by Steven F. Freeman entitled ***"The Unexplained Exit Poll Discrepancy."*** Freeman obtained exit poll data and the actual election results for 11 states that were considered by many to be the crucial states for the 2004 presidential election. The exit poll results show the number of voters polled as they left the voting booth for each state along with the corresponding percentage favoring Bush or Kerry and the predicted winner. The election results give the actual outcomes and winner for each state as reported by the state's election commission. The final column of the table shows the difference between the predicted winning percentage from the exit polls and the actual winning percentage from the election.

This table shows that the exit polls predicted George Bush to win in only 2 of the 11 crucial states, and this is why the media were predicting that John Kerry would win the election even before the polls were closed. In fact, Bush won 6 of the 11 crucial states, and, perhaps more importantly, we see in the final column that in 10 of these 11 states the difference between the actual margin of victory from the election results and the predicted margin of victory from the exit polls favored Bush.

At the end of this chapter, we will discuss some of the cautions one must take in using exit poll data to predict actual election outcomes.

2.2 Observational Studies

observational study

experimental study

explanatory variables
response variables

A study may be either observational or experimental. In an **observational study,** the researcher records information concerning the subjects under study without any interference with the process that is generating the information. The researcher is a passive observer of the transpiring events. In an **experimental study** (which will be discussed in detail in Sections 2.4 and 2.5), the researcher actively manipulates certain variables associated with the study, called the **explanatory variables,** and then records their effects on the **response variables** associated with the experimental subjects. A severe limitation of observational studies is that the recorded values

confounding variables

of the response variables may be affected by variables other than the explanatory variables. These variables are not under the control of the researcher. They are called **confounding variables.** The effects of the confounding variables and the explanatory variables on the response variable cannot be separated due to the lack of control the researcher has over the physical setting in which the observations are made. In an experimental study, the researcher attempts to maintain control over all variables that may have an effect on the response variables.

comparative study
descriptive study

Observational studies may be dichotomized into either a **comparative study** or a **descriptive study.** In a comparative study, two or more methods of achieving a result are compared for effectiveness. For example, three types of healthcare delivery methods are compared based on cost effectiveness. Alternatively, several groups are compared based on some common attribute. For example, the starting incomes of engineers are contrasted from a sample of new graduates from private and public universities. In a descriptive study, the major purpose is to characterize a population or process based on certain attributes in that population or process—for example, studying the health status of children under the age of 5 years old in families without health insurance or assessing the number of overcharges by companies hired under federal military contracts.

Observational studies in the form of polls, surveys, and epidemiological studies, for example, are used in many different settings to address questions posed by researchers. Surveys are used to measure the changing opinion of the nation with respect to issues such as gun control, interest rates, taxes, the minimum wage, Medicare, and the national debt. Similarly, we are informed on a daily basis through newspapers, magazines, television, radio, and the Internet of the results of public opinion polls concerning other relevant (and sometimes irrelevant) political, social, educational, financial, and health issues.

In an observational study, the factors (treatments) of interest are not manipulated while making measurements or observations. The researcher in an environmental impact study is attempting to establish the current state of a natural setting to which subsequent changes may be compared. Surveys are often used by natural scientists as well. In order to determine the proper catch limits of commercial and recreational fishermen in the Gulf of Mexico, the states along the Gulf of Mexico must sample the Gulf to determine the current fish density.

There are many biases and sampling problems that must be addressed in order for the survey to be a reliable indicator of the current state of the sampled population. A problem that may occur in observational studies is assigning **cause-and-effect relationships** to spurious associations between factors. For example, in many epidemiological studies, we study various environmental, social, and ethnic factors and their relationship with the incidence of certain diseases. A public health question of considerable interest is the relationship between heart disease and the amount of fat in one's diet. It would be unethical to randomly assign volunteers to one of several high-fat diets and then monitor the people over time to observe whether or not heart disease develops.

cause-and-effect relationships

Without being able to manipulate the factor of interest (fat content of the diet), the scientist must use an observational study to address the issue. This could be done by comparing the diets of a sample of people with heart disease with the diets of a sample of people without heart disease. Great care would have to be taken to record other relevant factors such as family history of heart disease, smoking habits, exercise routine, age, and gender for each person, along with other physical characteristics. Models could then be developed so that differences between the two groups could be adjusted to eliminate all factors except fat content of the diet.

Even with these adjustments, it would be difficult to assign a cause-and-effect relationship between the high fat content of a diet and the development of heart disease. In fact, if the dietary fat content for the heart disease group tended to be higher than that for the group free of heart disease after adjusting for relevant factors, the study results would be reported as an **association** between high dietary fat content and heart disease, not a **causal** relationship.

association
causal

Stated differently, in observational studies, we are sampling from populations where the factors (or treatments) are already present, and we compare samples with respect to the factors (treatments) of interest to the researcher. In contrast, in the controlled environment of an experimental study, we are able to randomly assign the people as objects under study to the factors (or treatments) and then observe the response of interest. For our heart disease example, the distinction is shown here:

Observational study: We sample from the heart disease population and heart disease–free population and compare the fat content of the diets for the two groups.

Experimental study: Ignoring ethical issues, we assign volunteers to one of several diets with different levels of dietary fat (the treatments) and compare the different treatments with respect to the response of interest (incidence of heart disease) after a period of time.

Observational studies are of three basic types:

sample survey

- A **sample survey** is a study that provides information about a population at a particular point in time (current information).

prospective study

- A **prospective study** is a study that observes a population in the present using a sample survey and proceeds to follow the subjects in the sample forward in time in order to record the occurrence of specific outcomes.

retrospective study

- A **retrospective study** is a study that observes a population in the present using a sample survey and also collects information about the subjects in the sample regarding the occurrence of specific outcomes that have already taken place.

In the health sciences, a sample survey would be referred to as a cross-sectional or prevalence study. All individuals in the survey would be asked about their current disease status and any past exposures to the disease. A prospective study would identify a group of disease-free subjects and then follow them over a period of time until some of the individuals develop the disease. The development or nondevelopment of the disease would then be related to other variables measured on the subjects at the beginning of the study, often referred to as exposure variables. A retrospective study identifies two groups of subjects: cases—subjects with the disease—and controls—subjects without the disease. The researcher then attempts to correlate the subjects' prior health habits to their current health status.

Although prospective and retrospective studies are both observational studies, there are some distinct differences.

- Retrospective studies are generally cheaper and can be completed more rapidly than prospective studies.
- Retrospective studies have problems due to inaccuracies in data due to recall errors.

- Retrospective studies have no control over variables that may affect disease occurrence.
- In prospective studies, subjects can keep careful records of their daily activities.
- In prospective studies, subjects can be instructed to avoid certain activities that may bias the study.
- Although prospective studies reduce some of the problems of retrospective studies, they are still observational studies, and hence the potential influences of confounding variables may not be completely controlled. It is possible to somewhat reduce the influence of the confounding variables by restricting the study to matched subgroups of subjects.

cohort studies
case-control studies

Both prospective and retrospective studies are often comparative in nature. Two specific types of such studies are **cohort studies** and **case-control studies.** In a cohort study, a group of subjects is followed forward in time to observe the differences in characteristics between subjects who develop a disease and those who do not. Similarly, we could observe which subjects commit crimes while also recording information about their educational and social backgrounds. In case-control studies, two groups of subjects are identified, one with the disease and one without the disease. Next, information is gathered about the subjects from their past concerning risk factors that are associated with the disease. Distinctions are then drawn between the two groups based on these characteristics.

EXAMPLE 2.1

A study was conducted to determine if women taking oral contraceptives had a greater propensity to develop heart disease. A group of 5,000 women currently using oral contraceptives and another group of 5,000 women not using oral contraceptives were selected for the study. At the beginning of the study, all 10,000 women were given physicals and were found to have healthy hearts. The women's health was then tracked for a 3-year period. At the end of the study, 15 of the 5,000 users had developed a heart disease, whereas only 3 of the nonusers had any evidence of heart disease. What type of design was this observational study?

Solution This study is an example of a prospective observational study. All women were free of heart disease at the beginning of the study and their exposure (oral contraceptive use) measured at that time. The women were then under observation for 3 years, with the onset of heart disease recorded if it occurred during the observation period. A comparison of the frequency of occurrence of the disease was made between the two groups of women, users and nonusers of oral contraceptives. ■

EXAMPLE 2.2

A study was designed to determine if people who use public transportation to travel to work are more politically active than people who use their own vehicle to travel to work. A sample of 100 people in a large urban city was selected from each group, and then all 200 individuals were interviewed concerning their political activities over the past 2 years. Out of the 100 people who used public transportation, 18 reported that they had actively assisted a candidate in the past 2 years, whereas only 9 of the 100 persons who used their own vehicles stated they had participated in a political campaign. What type of design was this study?

Solution This study is an example of a retrospective observational study. The individuals in both groups were interviewed about their past experiences with the political process. A comparison of the degree of participation of the individuals was made across the two groups. ■

In Example 2.2, many of the problems with using observational studies are present. There are many factors that may affect whether or not an individual decides to participate in a political campaign. Some of these factors may be confounded with ridership on public transportation—for example, awareness of the environmental impact of vehicular exhaust on air pollution, income level, and education level. These factors need to be taken into account when designing an observational study.

The most widely used observational study is the survey. Information from surveys impacts nearly every facet of our daily lives. Government agencies use surveys to make decisions about the economy and many social programs. News agencies often use opinion polls as a basis of news reports. Ratings of television shows, which come from surveys, determine which shows will be continued for the next television season.

Who conducts surveys? The various news organizations all use public opinion polls: Such surveys include the *New York Times/CBS News, Washington Post/ABC News, Wall Street Journal/NBC News, Harris, Gallup/Newsweek,* and *CNN/Time* polls. However, the vast majority of surveys are conducted for a specific industrial, governmental, administrative, political, or scientific purpose. For example, auto manufacturers use surveys to find out how satisfied customers are with their cars. Frequently, we are asked to complete a survey as part of the warranty registration process following the purchase of a new product. Many important studies involving health issues use surveys to determine, for example, the amount of fat in a diet, exposure to secondhand smoke, condom use and the prevention of AIDS, and the prevalence of adolescent depression.

The U.S. Bureau of the Census is required by the U.S. Constitution to enumerate the population every 10 years. With the growing involvement of the government in the lives of its citizens, the Census Bureau has expanded its role beyond just counting the population. An attempt is made to send a census questionnaire in the mail to every household in the United States. Since the 1940 census, in addition to the complete count information, further information has been obtained from representative samples of the population. In the 2000 census, variable sampling rates were employed. For most of the country, approximately five of six households were asked to answer the 14 questions on the short version of the form. The remaining households responded to a longer version of the form containing an additional 45 questions. Many agencies and individuals use the resulting information for many purposes. The federal government uses it to determine allocations of funds to states and cities. Businesses use it to forecast sales, to manage personnel, and to establish future site locations. Urban and regional planners use it to plan land use, transportation networks, and energy consumption. Social scientists use it to study economic conditions, racial balance, and other aspects of the quality of life.

The U.S. Bureau of Labor Statistics (BLS) routinely conducts more than 20 surveys. Some of the best known and most widely used are the surveys that establish the Consumer Price Index (CPI). The CPI is a measure of price change for a fixed market basket of goods and services over time. It is a measure of inflation and serves as an economic indicator for government policies. Businesses tie wage rates and pension plans to the CPI. Federal health and welfare programs, as well as many state and local programs, tie their bases of eligibility to the CPI.

Escalator clauses in rents and mortgages are based on the CPI. This one index, determined on the basis of sample surveys, plays a fundamental role in our society.

Many other surveys from the BLS are crucial to society. The monthly Current Population Survey establishes basic information on the labor force, employment, and unemployment. The Consumer Expenditure Survey collects data on family expenditures for goods and services used in day-to-day living. The Current Employment Statistics Survey collects information on employment hours and earnings for nonagricultural business establishments. The Occupational Employment Statistics Survey provides information on future employment opportunities for a variety of occupations, projecting to approximately 10 years ahead. Other activities of the BLS are addressed in the *BLS Handbook of Methods* (web version: www.bls.gov/opub/hom).

Opinion polls are constantly in the news, and the names of Gallup and Harris have become well known to everyone. These polls, or sample surveys, reflect the attitudes and opinions of citizens on everything from politics and religion to sports and entertainment. The Nielsen ratings determine the success or failure of TV shows.

How do you figure out the ratings? Nielsen Media Research (NMR) continually measures television viewing with a number of different samples all across the United States. The first step is to develop representative samples. This must be done with a scientifically drawn random selection process. No volunteers can be accepted or the statistical accuracy of the sample would be in jeopardy. Nationally, there are 5,000 television households in which electronic meters (called People Meters) are attached to every TV set, VCR, cable converter box, satellite dish, or other video equipment in the home. The meters continually record all set tunings. In addition, NMR asks each member of the household to let them know when they are watching by pressing a pre-assigned button on the People Meter. By matching this button activity to the demographic information (age/gender) NMR collected at the time the meters were installed, NMR can match the set tuning—what is being watched—with who is watching. All these data are transmitted to NMR's computers, where they are processed and released to customers each day. In addition to this national service, NMR has a slightly different metering system in 55 local markets. In each of those markets, NMR gathers just the set-tuning information each day from more than 20,000 additional homes. NMR then processes the data and releases what are called "household ratings" daily. In this case, the ratings report what channel or program is being watched, but they do not have the "who" part of the picture. To gather that local demographic information, NMR periodically (at least four times per year) asks another group of people to participate in diary surveys. For these estimates, NMR contacts approximately 1 million homes each year and asks them to keep track of television viewing for 1 week, recording their TV-viewing activity in a diary. This is done for all 210 television markets in the United States in November, February, May, and July and is generally referred to as the "sweeps." For more information on the Nielsen ratings, go to the NMR website (www. nielsenmedia. com) and click on the "What TV Ratings Really Mean" button.

Businesses conduct sample surveys for their internal operations in addition to using government surveys for crucial management decisions. Auditors estimate account balances and check on compliance with operating rules by sampling accounts. Quality control of manufacturing processes relies heavily on sampling techniques.

Another area of business activity that depends on detailed sampling activities is marketing. Decisions on which products to market, where to market them, and how to advertise them are often made on the basis of sample survey data. The data

may come from surveys conducted by the firm that manufactures the product or may be purchased from survey firms that specialize in marketing data.

2.3 Sampling Designs for Surveys

A crucial element in any survey is the manner in which the sample is selected from the population. If the individuals included in the survey are selected based on convenience alone, there may be biases in the sample survey, which would prevent the survey from accurately reflecting the population as a whole. For example, a marketing graduate student developed a new approach to advertising and, to evaluate this new approach, selected the students in a large undergraduate business course to assess whether the new approach is an improvement over standard advertisements. Would the opinions of this class of students be representative of the general population of people to which the new approach to advertising would be applied? The income levels, ethnicities, education levels, and many other socioeconomic characteristics of the students may differ greatly from the population of interest. Furthermore, the students may be coerced into participating in the study by their instructor and hence may not give the most candid answers to questions on a survey. Thus, the manner in which a sample is selected is of utmost importance to the credibility and applicability of the study's results.

In order to precisely describe the components that are necessary for a sample to be effective, the following definitions are required.

target population

Target population: The complete collection of objects whose description is the major goal of the study. Designating the target population is a crucial but often difficult part of the first step in an observational or experimental study. For example, in a survey to decide if a new storm-water drainage tax should be implemented, should the target population be all persons over the age of 18 in the county, all registered voters, or all persons paying property taxes? The selection of the target population may have a profound effect on the results of the study.

sample

Sample: A subset of the target population.

sampled population

Sampled population: The complete collection of objects that have the potential of being selected in the sample; the population from which the sample is *actually* selected. In many studies, the sampled population and the target population are very different. This may lead to very erroneous conclusions based on the information collected in the sample. For example, in a telephone survey of people who are on the property tax list (the target population), a subset of this population may not answer their telephone if the caller is unknown, as viewed through Caller ID. Thus, the sampled population may be quite different from the target population with respect to some important characteristics such as income and opinion on certain issues.

observation unit

Observation unit: The object about which data are collected. In studies involving human populations, the observation unit is a specific individual in the sampled population. In ecological studies, the observation unit may be a sample of water from a stream or an individual plant on a plot of land.

sampling unit

Sampling unit: The object that is actually sampled. We may want to sample the person who pays the property tax but may only have

a list of telephone numbers. Thus, the households in the sampled population serve as the sampled units, and the observation units are the individuals residing in the sampled household. In an entomology study, we may sample 1-acre plots of land and then count the number of insects on individual plants residing on the sampled plot. The sampled unit is the plot of land; the observation unit would be the individual plant.

sampling frame

Sampling frame: The list of sampling units. For a mailed survey, it may be a list of addresses of households in a city. For an ecological study, it may be a map of areas downstream from power plants.

In a perfect survey, the target population would be the same as the sampled population. This type of survey rarely happens. There are always difficulties in obtaining a sampling frame or being able to identify all elements within the target population. A particular aspect of this problem is nonresponse. Even if the researcher was able to obtain a list of all individuals in the target population, there may be a distinct subset of the target population that refuses to fill out the survey or allow themselves to be observed. Thus, the sampled population becomes a subset of the target population. An attempt at characterizing the nonresponders is very crucial in attempting to use a sample to describe a population. The group of nonresponders may have certain demographics or a particular political leaning that if not identified could greatly distort the results of the survey. An excellent discussion of this topic can be found in the textbook *Sampling: Design and Analysis* by Sharon L. Lohr (1999).

simple random sampling

The basic design (**simple random sampling**) consists of selecting a group of n units in such a way that each sample of size n has the same chance of being selected. Thus, we can obtain a random sample of eligible voters in a bond-issue poll by drawing names from the list of registered voters in such a way that each sample of size n has the same probability of selection. The details of simple random sampling are discussed in Section 4.11. At this point, we merely state that a simple random sample will contain as much information on community preference as any other sample survey design, provided all voters in the community have similar socioeconomic backgrounds.

Suppose, however, that the community consists of people in two distinct income brackets, high and low. Voters in the high-income bracket may have opinions on the bond issue that are quite different from the opinions of voters in the low-income bracket. Therefore, to obtain accurate information about the population, we want to sample voters from each bracket. We can divide the population elements into two groups, or strata, according to income and select a simple random sample from each group. The resulting sample is called a **stratified random sample.** (See Chapter 5 of Scheaffer et al., 2006.) Note that stratification is accomplished by using knowledge of an auxiliary variable, namely, personal income. By stratifying on high and low values of income, we increase the accuracy of our estimator. **Ratio estimation** is a second method for using the information contained in an auxiliary variable. Ratio estimators not only use measurements on the response of interest but also incorporate measurements on an auxiliary variable. Ratio estimation can also be used with stratified random sampling.

stratified random sample

ratio estimation

Although individual preferences are desired in the survey, a more economical procedure, especially in urban areas, may be to sample specific families, apartment buildings, or city blocks rather than individual voters. Individual preferences can then be obtained from each eligible voter within the unit sampled. This technique is called **cluster sampling.** Although we divide the population into groups

cluster sampling

for both cluster sampling and stratified random sampling, the techniques differ. In stratified random sampling, we take a simple random sample within each group, whereas in cluster sampling, we take a simple random sample of groups and then sample all items within the selected groups (clusters). (See Chapters 8 and 9 of Scheaffer et al., 2006, for details.)

Sometimes the names of persons in the population of interest are available in a list, such as a registration list, or on file cards stored in a drawer. For this situation, an economical technique is to draw the sample by selecting one name near the beginning of the list and then selecting every tenth or fifteenth name thereafter. If the sampling is conducted in this manner, we obtain a **systematic sample.** As you might expect, systematic sampling offers a convenient means of obtaining sample information; however, systematic sampling will be less precise than simple random sampling if the sampling frame has a periodicity. (Details are given in Chapter 7 of Scheaffer et al., 2006.)

systematic sample

The following example will illustrate how the goal of the study or the information available about the elements of the population determines which type of sampling design to use in a particular study.

EXAMPLE 2.3

Identify the type of sampling design in each of the following situations.

a. The selection of 200 people to serve as potential jurors in a medical malpractice trial is conducted by assigning a number to each of 140,000 registered voters in the county. A computer software program is used to randomly select 200 numbers from the numbers 1 to 140,000. The people having these 200 numbers are sent a postcard notifying them of their selection for jury duty.
b. Suppose you are selecting microchips from a production line for inspection for bent probes. As the chips proceed past the inspection point, every 100th chip is selected for inspection.
c. The Internal Revenue Service wants to estimate the amount of personal deductions taxpayers made based on the type of deduction: home office, state income tax, property taxes, property losses, and charitable contributions. The amount claimed in each of these categories varies greatly depending on the adjusted gross income of the taxpayer. Therefore, a simple random sample would not be an efficient design. The IRS decides to divide taxpayers into five groups based on their adjusted gross incomes and then takes a simple random sample of taxpayers from each of the five groups.
d. The USDA inspects produce for *E. coli* contamination. As trucks carrying produce cross the border, the truck is stopped for inspection. A random sample of five containers is selected for inspection from the hundreds of containers on the truck. Every apple in each of the five containers is then inspected for *E. coli.*

Solution

a. A simple random sample is selected using the list of registered voters as the sampling frame.
b. This is an example of systematic random sampling. This type of inspection should provide a representative sample of chips because there is no reason to presume that there exists any cyclic variation

in the production of the chips. It would be very difficult in this situation to perform simple random sampling because no sampling frame exists.

c. This is an example of stratified random sampling with the five levels of personal deductions serving as the strata. Overall the personal deductions of taxpayers increase with income. This results in the stratified random sample having a much smaller total sample size than would be required in a simple random sample to achieve the same level of precision in its estimators.

d. This is a cluster sampling design with the clusters being the containers and the individual apples being the measurement unit. ■

The important point to understand is that there are different kinds of surveys that can be used to collect sample data. For the surveys discussed in this text, we will deal with simple random sampling and methods for summarizing and analyzing data collected in such a manner. More complicated surveys lead to even more complicated problems at the summarization and analysis stages of statistics.

The American Statistical Association (http://www.amstat.org) publishes a booklet: *What Is a Survey?*. The booklet describes many of the elements crucial to obtaining a valid and useful survey. It lists many of the potential sources of errors commonly found in surveys with guidelines on how to avoid these pitfalls. A discussion of some of the issues raised in the booklet follows.

Problems Associated with Surveys

Even when the sample is selected properly, there may be uncertainty about whether the survey represents the population from which the sample was selected. Two of the major sources of uncertainty are nonresponse, which occurs when a portion of the individuals sampled cannot or will not participate in the survey, and measurement problems, which occur when the respondents' answers to questions do not provide the type of data that the survey was designed to obtain.

survey nonresponse

Survey nonresponse may result in a biased survey because the sample is not representative of the population. It is stated in *Judging the Quality of a Survey* that in surveys of the general population women are more likely to participate than men; that is, the nonresponse rate for males is higher than for females. Thus, a political poll may be biased if the percentage of women in the population in favor of a particular issue is larger than the percentage of men in the population supporting the issue. The poll would overestimate the percentage of the population in favor of the issue because the sample had a larger percentage of women than their percentage in the population. In all surveys, a careful examination of the nonresponse group must be conducted to determine whether a particular segment of the population may be either under- or overrepresented in the sample. Some of the remedies for nonresponse are

1. Offering an inducement for participating in the survey
2. Sending reminders or making follow-up telephone calls to the individuals who did not respond to the first contact
3. Using statistical techniques to adjust the survey findings to account for the sample profile differing from the population profile

measurement problems

Measurement problems are the result of the respondents not providing the information that the survey seeks. These problems often are due to the specific wording of questions in a survey, the manner in which the respondent answers the survey questions, and the fashion in which an interviewer phrases questions during the interview. Examples of specific problems and possible remedies are as follows:

1. *Inability to recall answers to questions:* The interviewee is asked how many times he or she visited a particular city park during the past year. This type of question often results in an underestimate of the average number of times a family visits the park during a year because people often tend to underestimate the number of occurrences of a common event or an event occurring far from the time of the interview. A possible remedy is to request respondents to use written records or to consult with other family members before responding.
2. *Leading questions:* The fashion in which an opinion question is posed may result in a response that does not truly represent the interviewee's opinion. Thus, the survey results may be biased in the direction in which the question is slanted. For example, a question concerning whether the state should impose a large fine on a chemical company for environmental violations is phrased as "Do you support the state fining the chemical company, which is the major employer of people in our community, considering that this fine may result in their moving to another state?" This type of question tends to elicit a "no" response and thus produces a distorted representation of the community's opinion on the imposition of the fine. The remedy is to write questions carefully in an objective fashion.
3. *Unclear wording of questions:* An exercise club attempted to determine the number of times a person exercises per week. The question asked of the respondent was "How many times in the last week did you exercise?" The word *exercise* has different meanings to different individuals. The result of allowing different definitions of important words or phrases in survey questions is to greatly reduce the accuracy of survey results. Several remedies are possible: The questions should be tested on a variety of individuals prior to conducting the survey to determine whether there are any confusing or misleading terms in the questions. During the training of the interviewers. all interviewers should have the "correct" definitions of all key words and be advised to provide these definitions to the respondents.

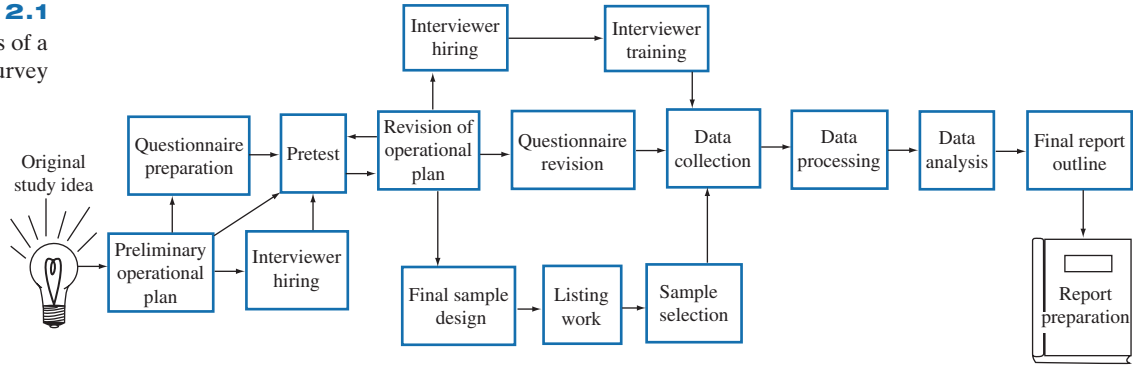
Many other issues, problems, and remedies are provided in the brochures from the ASA.

The stages in designing, conducting, and analyzing a survey are contained in Figure 2.1, which has been reproduced from an earlier version of *What Is a Survey?* in Cryer and Miller's *Statistics for Business: Data Analysis and Modeling* (1991). This diagram provides a guide for properly conducting a successful survey.

Data Collection Techniques

Having chosen a particular sample survey, how does one actually collect the data? The most commonly used methods of data collection in sample surveys are personal interviews and telephone interviews. These methods, with appropriately

FIGURE 2.1
Stages of a survey



trained interviewers and carefully planned callbacks, commonly achieve response rates of 60% to 75% and sometimes even higher. A mailed questionnaire sent to a specific group of interested persons can sometimes achieve good results, but generally the response rates for this type of data collection are so low that all reported results are suspect. Frequently, objective information can be found from direct observation rather than from an interview or mailed questionnaire.

personal interviews

Data are frequently obtained by **personal interviews.** For example, we can use personal interviews with eligible voters to obtain a sample of public sentiment toward a community bond issue. The procedure usually requires the interviewer to ask prepared questions and to record the respondent's answers. The primary advantage of these interviews is that people will usually respond when confronted in person. In addition, the interviewer can note specific reactions and eliminate misunderstandings about the questions asked. The major limitations of the personal interview (aside from the cost involved) concern the interviewers. If they are not thoroughly trained, they may deviate from the required protocol, thus introducing a bias into the sample data. Any movement, facial expression, or statement by the interviewer can affect the response obtained. For example, a leading question such as "Are you also in favor of the bond issue?" may tend to elicit a positive response. Finally, errors in recording the responses can lead to erroneous results.

telephone interviews

Information can also be obtained from persons in the sample through **telephone interviews.** With the competition among telephone service providers, an interviewer can place any number of calls to specified areas of the country relatively inexpensively. Surveys conducted through telephone interviews are frequently less expensive than personal interviews, owing to the elimination of travel expenses. The investigator can also monitor the interviews to be certain that the specified interview procedure is being followed.

A major problem with telephone surveys is that it is difficult to find a list or directory that closely corresponds to the population. Telephone directories have many numbers that do not belong to households, and many households have unlisted numbers. A technique that avoids the problem of unlisted numbers is random-digit dialing. In this method, a telephone exchange number (the first three digits of a seven-digit number) is selected, and then the last four digits are dialed randomly until a fixed number of households of a specified type are reached. This technique produces samples from the target population, but most random-digit-dialing samples include only landline numbers. Thus, the increasing number of households with cell phones only is excluded. Also, many people screen calls before answering a call. These two problems are creating potentially large biases in telephone surveys.

Telephone interviews generally must be kept shorter than personal interviews because responders tend to get impatient more easily when talking over the telephone. With appropriately designed questionnaires and trained interviewers, telephone interviews can be as successful as personal interviews.

self-administered questionnaire

Another useful method of data collection is the **self-administered questionnaire,** to be completed by the respondent. These questionnaires usually are mailed to the individuals included in the sample, although other distribution methods can be used. The questionnaire must be carefully constructed if it is to encourage participation by the respondents.

The self-administered questionnaire does not require interviewers, and thus its use results in savings in the survey cost. This savings in cost is usually bought at the expense of a lower response rate. Nonresponse can be a problem in any form of data collection, but since we have the least contact with respondents in a mailed questionnaire, we frequently have the lowest rate of response. The low response rate can introduce a bias into the sample because the people who answer questionnaires may not be representative of the population of interest. To eliminate some of the bias, investigators frequently contact the nonrespondents through follow-up letters, telephone interviews, or personal interviews.

direct observation

The fourth method for collecting data is **direct observation.** If we were interested in estimating the number of trucks that use a particular road during the 4–6 P.M. rush hours, we could assign a person to count the number of trucks passing a specified point during this period, or electronic counting equipment could be used. The disadvantage in using an observer is the possibility of error in observation.

Direct observation is used in many surveys that do not involve measurements on people. The USDA measures certain variables on crops in sections of fields in order to produce estimates of crop yields. Wildlife biologists may count animals, animal tracks, eggs, or nests to estimate the size of animal populations.

A closely related notion to direct observation is that of getting data from objective sources not affected by the respondents themselves. For example, health information can sometimes be obtained from hospital records and income information from employer's records (especially for state and federal government workers). This approach may take more time but can yield large rewards in important surveys.

2.4 Experimental Studies

An experimental study may be conducted in many different ways. In some studies, the researcher is interested in collecting information from an undisturbed natural process or setting. An example would be a study of the differences in reading scores of second-grade students in public, religious, and private schools. In other studies, the scientist is working within a highly controlled laboratory, a completely artificial setting for the study. For example, the study of the effect of humidity and temperature on the length of the life cycles of ticks would be conducted in a laboratory, since it would be impossible to control the humidity or temperature in the tick's natural environment. This control of the factors under study allows the entomologist to obtain results that can then be more easily attributed to differences in the levels of the temperature and humidity, since nearly all other conditions remain constant throughout the experiment. In a natural setting, many other factors are varying, and they may also result in changes in the life cycles of the ticks. However, the greater the control in these artificial settings, the less likely

the experiment is portraying the true state of nature. A careful balance between control of conditions and depiction of a reality must be maintained in order for the experiment to be useful. In this section and the next one, we will present some standard designs of experiments. In experimental studies, the researcher controls the crucial factors by one of two methods.

Method 1: The subjects in the experiment are randomly assigned to the treatments. For example, 10 rats are randomly assigned to each of the four dose levels of an experimental drug under investigation.

Method 2: Subjects are randomly selected from different populations of interest. For example, 50 male and 50 female dogs are randomly selected from animal shelters in large and small cities and tested for the presence of heartworms.

In Method 1, the researcher randomly selects experimental units from a homogeneous population of experimental units and then has complete control over the assignment of the units to the various treatments. In Method 2, the researcher has control over the random sampling from the treatment populations but not over the assignment of the experimental units to the treatments.

In experimental studies, it is crucial that the scientist follows a systematic plan established prior to running the experiment. The plan includes how all randomization is conducted, either the assignment of experimental units to treatments or the selection of units from the treatment populations. There may be extraneous factors present that may affect the experimental units. These factors may be present as subtle differences in the experimental units or slight differences in the surrounding environment during the conducting of the experiment. The randomization process ensures that, on the average, any large differences observed in the responses of the experimental units in different treatment groups can be attributed to the differences in the groups and not to factors that were not controlled during the experiment. The plan should also include many other aspects of how to conduct the experiment. Some of the items that should be included in such a plan are listed here:

1. The research objectives of the experiment
2. The selection of the factors that will be varied (the treatments)
3. The identification of extraneous factors that may be present in the experimental units or in the environment of the experimental setting (the blocking factors)
4. The characteristics to be measured on the experimental units (response variable)
5. The method of randomization, either randomly selecting experimental units from treatment populations or randomly assigning experimental units to treatments
6. The procedures to be used in recording the responses from the experimental units
7. The selection of the number of experimental units assigned to each treatment may require designating the level of significance and power of tests or the precision and reliability of confidence intervals
8. A complete listing of available resources and materials

Terminology

designed experiment

A **designed experiment** is an investigation in which a specified framework is provided in order to observe, measure, and evaluate groups with respect to a

designated response. The researcher controls the elements of the framework during the experiment in order to obtain data from which statistical inferences can provide valid comparisons of the groups of interest.

factors
measurements or observations

There are two types of variables in a experimental study. Controlled variables called **factors** are selected by the researcher for comparison. Response variables are **measurements** or **observations** that are recorded but not controlled by the researcher. The controlled variables form the comparison groups defined by the research hypothesis.

treatments

The **treatments** in an experimental study are the conditions constructed from the factors. The factors are selected by examining the questions raised by the research hypothesis. In some experiments, there may only be a single factor, and hence the treatments and levels of the factor would be the same. In most cases, we will have several factors, and the treatments are formed by combining levels of the factors. This type of **treatment design** is called a **factorial treatment design.**

treatment design
factorial treatment design

We will illustrate these ideas in the following example.

EXAMPLE 2.4

A researcher is studying the conditions under which commercially raised shrimp achieve maximum weight gain. Three water temperatures (25°, 30°, 35°) and four water salinity levels (10%, 20%, 30%, 40%) were selected for study. Shrimp were raised in containers with specified water temperatures and salinity levels. The weight gain of the shrimp in each container was recorded after a 6-week study period. There are many other factors that may affect weight gain, such as density of shrimp in the containers, variety of shrimp, size of shrimp, type of feeding, and so on. The experiment was conducted as follows: 24 containers were available for the study. A specific variety and size of shrimp was selected for study. The density of shrimp in the container was fixed at a given amount. One of the three water temperatures and one of the four salinity levels were randomly assigned to each of the 24 containers. All other identifiable conditions were specified to be maintained at the same level for all 24 containers for the duration of the study. In reality, there will be some variation in the levels of these variables. After 6 weeks in the tanks, the shrimp were harvested and weighed. Identify the response variable, factors, and treatments in this example.

Solution The response variable is weight of the shrimp at the end of the 6-week study. There are two factors: water temperature at three levels (25°, 30°, and 35°) and water salinity at four levels (10%, 20%, 30%, and 40%). We can thus create $3 \cdot 4 = 12$ treatments from the combination of levels of the two factors. These factor-level combinations representing the 12 treatments are shown here:

(25°, 10%)	(25°, 20%)	(25°, 30%)	(25°, 40%)
(30°, 10%)	(30°, 20%)	(30°, 30%)	(30°, 40%)
(35°, 10%)	(35°, 20%)	(35°, 30%)	(35°, 40%)

Following proper experimental procedures, 2 of the 24 containers would be randomly assigned to each of the 12 treatments. ■

In other circumstances, there may be a large number of factors, and hence the number of treatments may be so large that only a subset of all possible treatments would be examined in the experiment. For example, suppose we were investigating the effect of the following factors on the yield per acre of soybeans: Factor 1—Five Varieties of Soybeans, Factor 2—Three Planting Densities, Factor 3—Four

Levels of Fertilization, Factor 4—Six Locations Within Texas, and Factor 5—Three Irrigation Rates. From the five factors, we can form $5 \cdot 3 \cdot 4 \cdot 6 \cdot 3 = 1{,}080$ distinct treatments. This would make for a very large and expensive experiment. In this type of situation, a subset of the 1,080 possible treatments would be selected for studying the relationship between the five factors and the yield of soybeans. This type of experiment has a **fractional factorial treatment structure,** since only a fraction of the possible treatments are actually used in the experiment. A great deal of care must be taken in selecting which treatments should be used in the experiment so as to be able to answer as many of the researcher's questions as possible.

fractional factorial treatment structure

control treatment

A special treatment is called the **control treatment.** This treatment is the benchmark to which the effectiveness of each remaining treatment is compared. There are three situations in which a control treatment is particularly necessary. First, the conditions under which the experiments are conducted may prevent generally effective treatments from demonstrating their effectiveness. In this case, the control treatment consisting of *no treatment* may help to demonstrate that the experimental conditions are keeping the treatments from demonstrating the differences in their effectiveness. For example, an experiment is conducted to determine the most effective level of nitrogen in a garden growing tomatoes. If the soil used in the study has a high level of fertility prior to adding nitrogen to the soil, all levels of nitrogen will appear to be equally effective. However, if a treatment consisting of adding *no nitrogen*—the control—is used in the study, the high fertility of the soil will be revealed, since the control treatment will be just as effective as the nitrogen-added treatments.

A second type of control is the *standard method* treatment to which all other treatments are compared. In this situation, several new procedures are proposed to replace an already existing well-established procedure. A third type of control is the *placebo control.* In this situation, a response may be obtained from the subject just by the manipulation of the subject during the experiment. A person may demonstrate a temporary reduction in pain level just by visiting with the physician and having a treatment prescribed. Thus, in evaluating several different methods of reducing pain level in patients, a treatment with no active ingredients, the placebo, is given to a set of patients without the patients' knowledge. The treatments with active ingredients are then compared to the placebo to determine their true effectiveness.

experimental unit

The **experimental unit** is the physical entity to which the treatment is randomly assigned or the subject that is randomly selected from one of the treatment populations. For the shrimp study of Example 2.4, the experimental unit is the container.

Consider another experiment in which a researcher is testing various dose levels (treatments) of a new drug on laboratory rats. If the researcher randomly assigned a single dose of the drug to each rat, then the experimental unit would be the individual rat. Once the treatment is assigned to an experimental unit, a single

replication

replication of the treatment has occurred. In general, we will randomly assign several experimental units to each treatment. We will thus obtain several independent observations on any particular treatment and hence will have several replications of the treatments. In Example 2.4, we had two replications of each treatment.

measurement unit

Distinct from the experimental unit is the **measurement unit.** This is the physical entity upon which a measurement is taken. In many experiments, the experimental and measurement units are identical. In Example 2.4, the measurement unit is the container, the same as the experimental unit. However, if the individual shrimp were weighed as opposed to obtaining the total weight of all the shrimp in

each container, the experimental unit would be the container, but the measurement unit would be the individual shrimp.

EXAMPLE 2.5

Consider the following experiment. Four types of protective coatings for frying pans are to be evaluated. Five frying pans are randomly assigned to each of the four coatings. The abrasion resistance of the coating is measured at three locations on each of the 20 pans. Identify the following items for this study: experimental design, treatments, replications, experimental unit, measurement unit, and total number of measurements.

Solution

Experimental design: Completely randomized design.
Treatments: Four types of protective coatings.
Replication: There are five frying pans (replications) for each treatment.
Experimental unit: Frying pan, because coatings (treatments) are randomly assigned to the frying pans.
Measurement unit: Particular locations on the frying pan.
Total number of measurements: $4 \cdot 5 \cdot 3 = 60$ measurements in this experiment.
The experimental unit is the frying pan, since the treatment was randomly assigned to a coating. The measurement unit is a location on the frying pan. ■

experimental error

The term **experimental error** is used to describe the variation in the responses among experimental units that are assigned the same treatment and are observed under the same experimental conditions. The reasons that the experimental error is not zero include (a) the natural differences in the experimental units prior to their receiving the treatment, (b) the variation in the devices that record the measurements, (c) the variation in setting the treatment conditions, and (d) the effect on the response variable of all extraneous factors other than the treatment factors.

EXAMPLE 2.6

Refer to the previously discussed laboratory experiment in which the researcher randomly assigns a single dose of the drug to each of 10 rats and then measures the level of the drug in the rats' bloodstream after 2 hours. For this experiment, the experimental unit and measurement unit are the same: the rat.

Identify the four possible sources of experimental error for this study. (See (a) to (d) in the last paragraph before this example.)

Solution We can address these sources as follows:

a. Natural differences in experimental units prior to receiving the treatment. There will be slight physiological differences among rats, so two rats receiving the exact same dose level (treatment) will have slightly different blood levels 2 hours after receiving the treatment.
b. Variation in the devices used to record the measurements. There will be differences in the responses due to the method by which the quantity of the drug in the rat is determined by the laboratory technician. If several determinations of drug level were made in the

blood of the same rat, there may be differences in the amount of drug found due to equipment variation, technician variation, or conditions in the laboratory.

c. Variation in setting the treatment conditions. If there is more than one replication per treatment, the treatment may not be exactly the same from one rat to another. Suppose, for example, that we had 10 replications of each dose (treatment). It is highly unlikely that each of the 10 rats would receive exactly the same dose of drug specified by the treatment. There could be slightly different amounts of the drug in the syringes, and slightly different amounts could be injected and enter the bloodstreams.

d. The effect on the response variable (blood level) of all extraneous factors other than the treatment factors. Presumably, the rats are all placed in cages and given the same amount of food and water prior to determining the amount of the drug in their blood. However, the temperature, humidity, external stimulation, and other conditions may be somewhat different in the 10 cages. This may have an effect on the responses of the 10 rats.

Thus, these differences and variation in the external conditions within the laboratory during the experiment all contribute to the size of the experimental error in the experiment. ■

EXAMPLE 2.7

Refer to Example 2.4. Suppose that each treatment is assigned to two containers and that 40 shrimp are placed in each container. After 6 weeks, the individual shrimp are weighed. Identify the experimental units, measurement units, factors, treatments, number of replications, and possible sources of experimental error.

Solution This is a factorial treatment design with two factors: temperature and salinity level. The treatments are constructed by selecting a temperature and salinity level to be assigned to a particular container. We would have a total of $3 \cdot 4 = 12$ possible treatments for this experiment. The 12 treatments are

(25°, 10%)	(25°, 20%)	(25°, 30%)	(25°, 40%)
(30°, 10%)	(30°, 20%)	(30°, 30%)	(30°, 40%)
(35°, 10%)	(35°, 20%)	(35°, 30%)	(35°, 40%)

We next randomly assign two containers to each of the 12 treatments. This results in two replications of each treatment. The experimental unit is the container, since the individual containers are randomly assigned to a treatment. Forty shrimp are placed in the containers, and after 6 weeks, the weights of the individual shrimp are recorded. The measurement unit is the individual shrimp, since this is the physical entity upon which an observation is made. Thus, in this experiment the experimental and measurement units are different. Several possible sources of experimental error include the difference in the weights of the shrimp prior to being placed in the container, how accurately the temperature and salinity levels are maintained over the 6-week study period, how accurately the shrimp are weighed at the conclusion of the study, the consistency of the amount of food fed to the shrimp (whether each shrimp was given exactly the same quantity of food over the 6 weeks), and the variation in any other conditions that may affect shrimp growth. ■

2.5 Designs for Experimental Studies

The subject of designs for experimental studies cannot be given much justice at the beginning of a statistical methods course—entire courses at the undergraduate and graduate levels are needed to get a comprehensive understanding of the methods and concepts of experimental design. Even so, we will attempt to give you a brief overview of the subject because much of the data requiring summarization and analysis arises from experimental studies involving one of a number of designs. We will work by way of examples.

A consumer testing agency decides to evaluate the wear characteristics of four major brands of tires. For this study, the agency selects four cars of a standard car model and four tires of each brand. The tires will be placed on the cars and then driven 30,000 miles on a 2-mile racetrack. The decrease in tread thickness over the 30,000 miles is the variable of interest in this study. Four different drivers will drive the cars, but the drivers are professional drivers with comparable training and experience. The weather conditions, smoothness of the track, and the maintenance of the four cars will be essentially the same for all four brands over the study period. All extraneous factors that may affect the tires are nearly the same for all four brands. Thus, the testing agency feels confident that if there is a difference in wear characteristics between the brands at the end of the study, then this is truly a difference in the four brands and not a difference due to the manner in which the study was conducted. The testing agency is interested in recording other factors, such as the cost of the tires, the length of warranty offered by the manufacturer, whether the tires go out of balance during the study, and the evenness of wear across the width of the tires. In this example, we will consider only tread wear. There should be a recorded tread wear for each of the 16 tires, 4 tires for each brand. The methods presented in Chapters 8 and 15 could be used to summarize and analyze the sample tread-wear data in order to make comparisons (inferences) among the four tire brands. One possible inference of interest could be the selection of the brand having minimum tread wear. Can the best-performing tire brand in the sample data be expected to provide the best tread wear if the same study is repeated? Are the results of the study applicable to the driving habits of the typical motorist?

Experimental Designs

There are many ways in which the tires can be assigned to the four cars. We will consider one running of the experiment in which we have four tires of each of the four brands. First, we need to decide how to assign the tires to the cars. We could randomly assign a single brand to each car, but this would result in a design having as the unit of measurement the total loss of tread for all four tires on the car and not the individual tire loss. Thus, we must randomly assign the 16 tires to the four cars. In Chapter 15, we will demonstrate how this randomization is conducted. One possible arrangement of the tires on the cars is shown in Table 2.2.

completely randomized design

In general, a **completely randomized design** is used when we are interested in comparing t "treatments" (in our case, $t = 4$; the treatments are the tire brands). For each of the treatments, we obtain a sample of observations. The sample sizes could be different for the individual treatments. For example, we could test 20 tires from Brands A, B, and C but only 12 tires from Brand D. The sample of observations from a treatment is assumed to be the result of a simple random sample of observations from the hypothetical population of possible values that could have

TABLE 2.2
Completely randomized design of tire wear

Car 1	Car 2	Car 3	Car 4
Brand B	Brand A	Brand A	Brand D
Brand B	Brand A	Brand B	Brand D
Brand B	Brand C	Brand C	Brand D
Brand C	Brand C	Brand A	Brand D

resulted from that treatment. In our example, the sample of four tire-wear thicknesses from Brand A was considered to be the outcome of a simple random sample of four observations selected from the hypothetical population of possible tire-wear thicknesses for standard model cars traveling 30,000 miles using Brand A.

The experimental design could be altered to accommodate the effect of a variable related to how the experiment is conducted. In our example, we assumed that the effect of the different cars, weather, drivers, and various other factors was the same for all four brands. Now, if the wear on tires imposed by Car 4 was less severe than that of the other three cars, would our design take this effect into account? Because Car 4 had all four tires of Brand D placed on it, the wear observed for Brand D may be less than the wear observed for the other three brands because all four tires of Brand D were on the "best" car. In some situations, the objects being observed have existing differences prior to their assignment to the treatments. For example, in an experiment evaluating the effectiveness of several drugs for reducing blood pressure, the age or physical condition of the participants in the study may decrease the effectiveness of the drugs. To avoid masking the effectiveness of the drugs, we would want to take these factors into account. Also, the environmental conditions encountered during the experiment may reduce the effectiveness of the treatment.

randomized block design

In our example, we would want to avoid having the comparison of the tire brands distorted by the differences in the four cars. The experimental design used to accomplish this goal is called a **randomized block design** because we want to "block" out any differences in the four cars to obtain a precise comparison of the four brands of tires. In a randomized block design, each treatment appears in every block. In the blood pressure example, we would group the patients according to the severity of their blood pressure problem and then randomly assign the drugs to the patients within each group. Thus, the randomized block design is similar to a stratified random sample used in surveys. In the tire-wear example, we would use the four cars as the blocks and randomly assign one tire of each brand to each of the four cars, as shown in Table 2.3. Now, if there are any differences in the cars that may affect tire wear, that effect will be equally applied to all four brands.

What happens if the position of the tires on the car affects the wear on the tire? The positions on the car are right front (RF), left front (LF), right rear (RR), and left rear (LR). In Table 2.3, suppose that all four tires from Brand A are placed on the RF position, Brand B on RR, Brand C on LF, and Brand D on LR. Now,

TABLE 2.3
Randomized block design of tire wear

Car 1	Car 2	Car 3	Car 4
Brand A	Brand A	Brand A	Brand A
Brand B	Brand B	Brand B	Brand B
Brand C	Brand C	Brand C	Brand C
Brand D	Brand D	Brand D	Brand D

TABLE 2.4
Latin square design of tire wear

Position	Car 1	Car 2	Car 3	Car 4
RF	Brand A	Brand B	Brand C	Brand D
RR	Brand B	Brand C	Brand D	Brand A
LF	Brand C	Brand D	Brand A	Brand B
LR	Brand D	Brand A	Brand B	Brand C

if the greatest wear occurs for tires placed on the RF, then Brand A would be at a great disadvantage when compared to the other three brands. In this type of situation, we would state that the effect of brand and the effect of position on the car were confounded; that is, using the data in the study, the effects of two or more factors cannot be unambiguously attributed to a single factor. If we observed a large difference in the average wear among the four brands, is this difference due to differences in the brands or differences due to the position of the tires on the car? Using the design given in Table 2.3, this question cannot be answered. Thus, we now need two blocking variables: the "car" the tire is placed on and the "position" on the car. A design having two blocking variables is called a **Latin square design.** A Latin square design for our example is shown in Table 2.4.

Latin square design

Note that with this design, each brand is placed in each of the four positions and on each of the four cars. Thus, if position or car has an effect on the wear of the tires, the position effect and/or car effect will be equalized across the four brands. The observed differences in wear can now be attributed to differences in the brand of the tire.

The randomized block and Latin square designs are both extensions of the completely randomized design in which the objective is to compare t treatments. The analysis of data for a completely randomized design and for block designs and the inferences made from such analyses are discussed further in Chapters 14, 15, and 17. A special case of the randomized block design is presented in Chapter 6, where the number of treatments is $t = 2$ and the analysis of data and the inferences from these analyses are discussed.

Factorial Treatment Structure in a Completely Randomized Design

factors

In this section, we will discuss how treatments are constructed from several **factors** rather than just being t levels of a single factor. These types of experiments are involved with examining the effect of two or more independent variables on a response variable y. For example, suppose a company has developed a new adhesive for use in the home and wants to examine the effects of temperature and humidity on the bonding strength of the adhesive. Several treatment design questions arise in any study. First, we must consider what factors (independent variables) are of greatest interest. Second, the number of levels and the actual settings of these levels must be determined for each factor. Third, having separately selected the levels for each factor, we must choose the factor-level combinations (treatments) that will be applied to the experimental units.

The ability to choose the factors and the appropriate settings for each of the factors depends on the budget, the time to complete the study, and, most important, the experimenter's knowledge of the physical situation under study. In many cases, this will involve conducting a detailed literature review to determine the current state of knowledge in the area of interest. Then, assuming that the

experimenter has chosen the levels of each independent variable, he or she must decide which factor-level combinations are of greatest interest and are viable. In some situations, certain factor-level combinations will not produce an experimental setting that can elicit a reasonable response from the experimental unit. Certain combinations may not be feasible due to toxicity or practicality issues.

one-at-a-time approach

One approach for examining the effects of two or more factors on a response is called the **one-at-a-time approach.** To examine the effect of a single variable, an experimenter varies the levels of this variable while holding the levels of the other independent variables fixed. This process is continued until the effect of each variable on the response has been examined.

For example, suppose we want to determine the combination of nitrogen and phosphorus that produces the maximum amount of corn per plot. We would select a level of phosphorus (say, 20 pounds), vary the levels of nitrogen, and observe which combination gives maximum yield in terms of bushels of corn per acre. Next, we would use the level of nitrogen producing the maximum yield, vary the amount of phosphorus, and observe the combination of nitrogen and phosphorus that produces the maximum yield. This combination would be declared the "best" treatment. The problem with this approach will be illustrated using the hypothetical yield values given in Table 2.5. These values would be unknown to the experimenter. We will assume that many replications of the treatments are used in the experiment so that the experimental results are nearly the same as the true yields.

Initially, we run experiments with 20 pounds of phosphorus and the levels of nitrogen at 40, 50, and 60. We would determine that using 60 pounds of nitrogen with 20 pounds of phosphorus produces the maximum production, 160 bushels per acre. Next, we set the nitrogen level at 60 pounds and vary the phosphorus levels. This would result in the 10 level of phosphorus producing the highest yield, 175 bushels, when combined with 60 pounds of nitrogen. Thus, we would determine that 10 pounds of phosphorus with 60 pounds of nitrogen produces the maximum yield. The results of these experiments are summarized in Table 2.6.

Based on the experimental results using the one-factor-at-a-time methodology, we would conclude that the 60 pounds of nitrogen and 10 pounds of phosphorus is the optimal combination. An examination of the yields in Table 2.5 reveals that the true optimal combination was 40 pounds of nitrogen with 30 pounds of phosphorus, producing a yield of 190 bushels per acre. Thus, this type of experimentation may produce incorrect results whenever the effect of one factor on the response does not remain the same at all levels of the second factor. In this

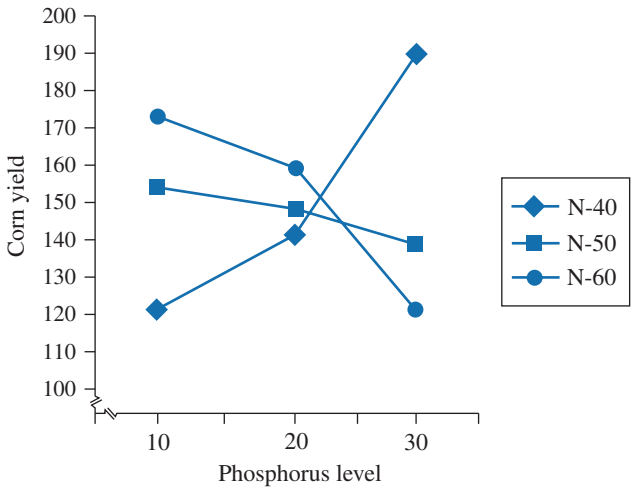
TABLE 2.5
Hypothetical population yields (bushels per acre)

	Phosphorus		
Nitrogen	**10**	**20**	**30**
40	125	145	190
50	155	150	140
60	175	160	125

TABLE 2.6
Yields for the experimental results

Phosphorus	20	20	20	10	30
Nitrogen	40	50	60	60	60
Yield	145	150	160	175	125

FIGURE 2.2
Yields from nitrogen–phosphorus treatments (interaction is present)



interact

situation, the factors are said to **interact.** Figure 2.2 depicts the interaction between nitrogen and phosphorus in the production of corn. Note that as the amount of nitrogen is increased from 40 to 60, there is an increase in the yield when using the 10 level of phosphorus. At the 20 level of phosphorus, increasing the amount of nitrogen also produces an increase in the yield but with smaller increments. At the 20 level of phosphorus, the yield increases 15 bushels when the nitrogen level is changed from 40 to 60. However, at the 10 level of phosphorus, the yield increases 50 bushels when the level of nitrogen is increased from 40 to 60. Furthermore, at the 30 level of phosphorus, increasing the level of nitrogen actually causes the yield to decrease. When there is no interaction between the factors, increasing the nitrogen level would have produced identical changes in the yield at all levels of phosphorus.

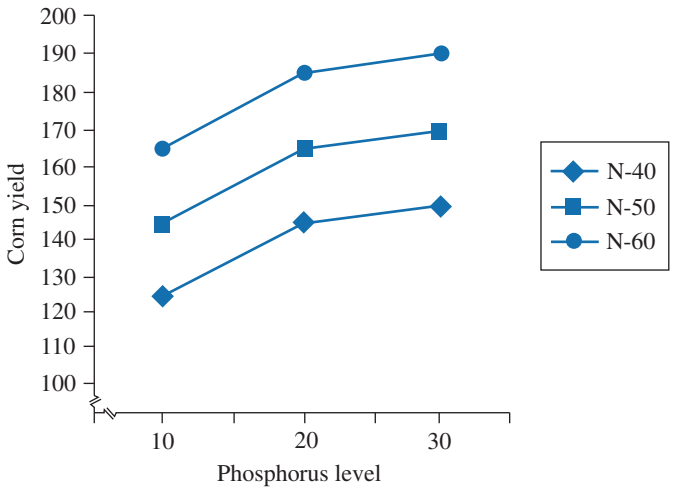
Table 2.7 and Figure 2.3 depict a situation in which the two factors do not interact. In this situation, the effect of phosphorus on the corn yield is the same for all three levels of nitrogen; that is, as we increase the amount of phosphorus, the change in corn yield is exactly the same for all three levels of nitrogen. Note that the change in yield is the same at all levels of nitrogen for a given change in phosphorus. However, the yields are larger at the higher levels of nitrogen. Thus, in the profile plots we have three different lines, but the lines are parallel. When interaction exists among the factors, the lines will either cross or diverge.

From Figure 2.3, we can observe that the one-at-a-time approach is appropriate for a situation in which the two factors do not interact. No matter what level is selected for the initial level of phosphorus, the one-at-a-time approach will produce the optimal yield. However, in most situations, prior to running the experiments it is not known whether the two factors will interact. If it is assumed that the factors *do not* interact and the one-at-a-time approach is implemented when in fact

TABLE 2.7
Hypothetical population yields (no interaction)

	Phosphorus		
Nitrogen	**10**	**20**	**30**
40	125	145	150
50	145	165	170
60	165	185	190

FIGURE 2.3
Yields from nitrogen–phosphorus treatments (no interaction)



the factors *do* interact, the experiment will produce results that will often fail to identify the best treatment.

factorial treatment structures

Factorial treatment structures are useful for examining the effects of two or more factors on a response, whether or not interaction exists. As before, the choice of the number of levels of each variable and the actual settings of these variables is important. When the factor-level combinations are assigned to experimental units at random, we have a completely randomized design with treatments being the factor-level combinations.

Using our previous example, we are interested in examining the effect of nitrogen and phosphorus levels on the yield of a corn crop. The nitrogen levels are 40, 50, and 60 pounds per plot, and the phosphorus levels are 10, 20, and 30 pounds per plot. We could use a completely randomized design where the nine factor-level combinations (treatments) of Table 2.8 are assigned at random to the experimental units (the plots of land planted with corn).

It is not necessary to have the same number of levels of both factors. For example, we could run an experiment with two levels of phosphorus and three levels of nitrogen, a 2×3 factorial structure. Also, the number of factors can be more than two. The corn yield experiment could have involved treatments consisting of four levels of potassium along with the three levels of phosphorus and nitrogen, a $4 \times 3 \times 3$ factorial structure. Thus, we would have $4 \cdot 3 \cdot 3 = 36$ factor combinations or treatments. The methodology of randomization, analysis, and inferences for data obtained from factorial treatment structures in various experimental designs is discussed in Chapters 14, 15, 17, and 18.

More Complicated Designs

Sometimes the objectives of a study are such that we wish to investigate the effects of certain factors on a response while blocking out certain other extraneous

TABLE 2.8
Factor-level combinations for the 3×3 factorial treatment structure

Treatment	1	2	3	4	5	6	7	8	9
Phosphorus	10	10	10	20	20	20	30	30	30
Nitrogen	40	50	60	40	50	60	40	50	60

TABLE 2.9
Block design for heartworm experiment

	Litter			
Puppy	**1**	**2**	**3**	**4**
1	A-D1	A-D3	B-D3	B-D2
2	A-D3	B-D1	A-D2	A-D2
3	B-D1	A-D1	B-D2	A-D1
4	A-D2	B-D2	B-D1	B-D3
5	B-D3	B-D3	A-D1	A-D3
6	B-D2	A-D2	A-D3	B-D1

sources of variability. Such situations require a block design with treatments from a factorial treatment structure and can be illustrated with the following example.

An investigator wants to examine the effectiveness of two drugs (A and B) for controlling heartworms in puppies. Veterinarians have conjectured that the effectiveness of the drugs may depend on a puppy's diet. Three different diets (Factor 1) are combined with the two drugs (Factor 2), and we have a 3×2 factorial treatment structure consisting of six treatments. Also, the effectiveness of the drugs may depend on a transmitted inherent protection against heartworms obtained from the puppy's mother. Thus, four litters of puppies consisting of six puppies each were selected to serve as a blocking factor in the experiment because all puppies within a given litter have the same mother. The six factor-level combinations (treatments) were randomly assigned to the six puppies within each of the four litters. The design is shown in Table 2.9. Note that this design is really a randomized **block design** in which the blocks are litters and the treatments are the six factor-level combinations of the 3×2 factorial treatment structure.

block design

Other more complicated combinations of block designs and factorial treatment structures are possible. As with sample surveys, however, we will deal only with the simplest experimental designs in this text. The point we want to make is that there are many different experimental designs that can be used in scientific studies for designating the collection of sample data. Each has certain advantages and disadvantages. We expand our discussion of experimental designs in Chapters 14–18, where we concentrate on the analysis of data generated from these designs. In those situations that require more complex designs, a professional statistician needs to be consulted to obtain the most appropriate design for the survey or experimental setting.

Controlling Experimental Error

As we observed in Examples 2.4 and 2.5, there are many potential sources of experimental error in an experiment. When the variance of experimental errors is large, the precision of our inferences will be greatly compromised. Thus, any techniques that can be implemented to reduce experimental error will lead to a much improved experiment and more precise inferences.

The researcher may be able to control many of the potential sources of experimental errors. Some of these sources are (1) the procedures under which the experiment is conducted, (2) the choice of experimental units and measurement units, (3) the procedure by which measurements are taken and recorded, (4) the blocking of the experimental units, (5) the type of experimental design, and (6)

covariates

the use of ancillary variables (called **covariates**). We will now address how each of these sources may affect experimental error and how the researcher may minimize the effect of these sources on the size of the variance of experimental error.

Experimental Procedures

When the individual procedures required to conduct an experiment are not followed in a careful, precise manner, the result is an increase in the variance of the response variable. This involves not only the personnel used to conduct the experiments and to measure the response variable but also the equipment used in their procedures. Personnel must be trained properly in constructing the treatments and carrying out the experiments. The consequences of their performance for the success of the experiment should be emphasized. The researcher needs to provide the technicians with equipment that will produce the most precise measurements within budget constraints. It is crucial that equipment be maintained and calibrated at frequent intervals throughout the experiment. The conditions under which the experiments are run must be as nearly constant as possible during the duration of the experiment. Otherwise, differences in the responses may be due to changes in the experimental conditions and not due to treatment differences.

When experimental procedures are not of high quality, the variance of the response variable may be inflated. Improper techniques used when taking measurements, improper calibration of instruments, or uncontrolled conditions within a laboratory may result in extreme observations that are not truly reflective of the effect of the treatment on the response variable. Extreme observations may also occur due to recording errors by the laboratory technician or the data manager. In either case, the researcher must investigate the circumstances surrounding extreme observations and then decide whether to delete the observations from the analysis. If an observation is deleted, an explanation of why the data value was not included should be given in the appendix of the final report.

When experimental procedures are not uniformly conducted throughout the study period, two possible outcomes are an inflation in the variance of the response variable and a bias in the estimation of the treatment mean. For example, suppose we are measuring the amount of a drug in the blood of rats injected with one of four possible doses of the drug. The equipment used to measure the precise amount of the drug to be injected is not working properly. For a given dosage of the drug, the first rats injected were given a dose that was less than the prescribed dose, whereas the last rats injected were given more than the prescribed amount. Thus, when the amount of the drug in the blood is measured, there will be an increase in the variance in these measurements, but the treatment mean may be estimated without bias because the overdose and underdose may cancel each other. On the other hand, if all the rats receiving the lowest dose level are given too much of the drug and all the rats receiving the highest dose level are not given enough of the drug, then the estimation of the treatment means will be biased. The treatment mean for the low dose will be overestimated, whereas the high dose will have an underestimated treatment mean. Thus, it is crucial to the success of the study that experimental procedures are conducted uniformly across all experimental units. The same is true concerning the environmental conditions within a laboratory or in a field study. Extraneous factors such as temperature, humidity, amount of sunlight, exposure to pollutants in the air, and other uncontrolled factors when not uniformly applied to the experimental units may result in a study with both an inflated variance and a biased estimation of treatment means.

Selecting Experimental and Measurement Units

When the experimental units used in an experiment are not similar with respect to those characteristics that may affect the response variable, the experimental error variance will be inflated. One of the goals of a study is to determine whether there is a difference in the mean responses of experimental units receiving different treatments. The researcher must determine the population of experimental units that are of interest. The experimental units are randomly selected from that population and then randomly assigned to the treatments. This is of course the idealized situation. In practice, the researcher is somewhat limited in the selection of experimental units by cost, availability, and ethical considerations. Thus, the inferences that can be drawn from the experimental data may be somewhat restricted. When examining the pool of potential experimental units, sets of units that are more similar in characteristics will yield more precise comparisons of the treatment means. However, if the experimental units are overly uniform, then the population to which inferences may be properly made will be greatly restricted. Consider the following example.

EXAMPLE 2.8

A sales campaign to market children's products will use television commercials as its central marketing technique. A marketing firm hired to determine whether the attention span of children is different depending on the type of product being advertised decided to examine four types of products: sporting equipment, healthy snacks, shoes, and video games. The firm selected 100 fourth-grade students from a New York City public school to participate in the study. Twenty-five students were randomly assigned to view a commercial for each of the four types of products. The attention spans of the 100 children were then recorded. The marketing firm thought that by selecting participants of the same grade level and from the same school system it would achieve a homogeneous group of subjects. What problems exist with this selection procedure?

Solution The marketing firm was probably correct in assuming that by selecting the students from the same grade level and school system it would achieve a more homogeneous set of experimental units than by using a more general selection procedure. However, this procedure has severely limited the inferences that can be made from the study. The results may be relevant only to students in the fourth grade and residing in a very large city. A selection procedure involving other grade levels and children from smaller cities would provide a more realistic study. ■

Reducing Experimental Error Through Blocking

When we are concerned that the pool of available experimental units has large differences with respect to important characteristics, the use of blocking may prove to be highly effective in reducing the experimental error variance. The experimental units are placed into groups based on their similarity with respect to characteristics that may affect the response variable. This results in sets or blocks of experimental units that are homogeneous within the block, but there is a broad coverage of important characteristics when considering the entire unit. The treatments are randomly assigned separately within each block. The comparison of the treatments is within the groups of homogeneous units and hence yields a comparison of the treatments that is not masked by the large differences in the original set of experimental

units. The blocking design will enable us to separate the variability associated with the characteristics used to block the units from the experimental error.

There are many criteria used to group experimental units into blocks; they include the following:

1. Physical characteristics such as age, weight, sex, health, and education of the subjects
2. Units that are related such as twins or animals from the same litter
3. Spatial location of experimental units such as neighboring plots of land or position of plants on a laboratory table
4. Time at which experiment is conducted such as the day of the week, because the environmental conditions may change from day to day
5. Person conducting the experiment, because if several operators or technicians are involved in the experiment, they may have some differences in how they make measurements or manipulate the experimental units

In all of these examples, we are attempting to observe all the treatments at each of the levels of the blocking criterion. Thus, if we were studying the number of cars with a major defect coming off each of three assembly lines, we might want to use day of the week as a blocking variable and be certain to compare each of the assembly lines on all 5 days of the work week.

Using Covariates to Reduce Variability

A covariate is a variable that is related to the response variable. Physical characteristics of the experimental units are used to create blocks of homogeneous units. For example, in a study to compare the effectiveness of a new diet to that of a control diet in reducing the weight of dogs, suppose the pool of dogs available for the study varied in age from 1 year to 12 years. We could group the dogs into three blocks: B_1—under 3 years, B_2—3 years to 8 years, B_3—over 8 years. A more exacting methodology records the age of the dog and then incorporates the age directly into the model when attempting to assess the effectiveness of the diet. The response variable would be adjusted for the age of the dog prior to comparing the new diet to the control diet. Thus, we have a more exact comparison of the diets. Instead of using a range of ages as is done in blocking, we are using the exact age of the dog, which reduces the variance of the experimental error.

Candidates for covariates in a given experiment depend on the particular experiment. The covariate needs to have a relationship to the response variable, it must be measurable, and it cannot be affected by the treatment. In most cases, the covariate is measured on the experimental unit before the treatment is given to the unit. Examples of covariates are soil fertility, amount of impurity in a raw material, weight of an experimental unit, SAT score of a student, cholesterol level of a subject, and insect density in a field. The following example will illustrate the use of a covariate.

EXAMPLE 2.9

In this study, the effects of two treatments, supplemental lighting (S) and partial shading (P), on the yield of soybean plants were compared with normal lighting (C). Normal lighting will serve as a control. Each type of lighting was randomly assigned to 15 soybean plants, and the plants were grown in a greenhouse study. When setting up the experiment, the researcher recognized that the plants were

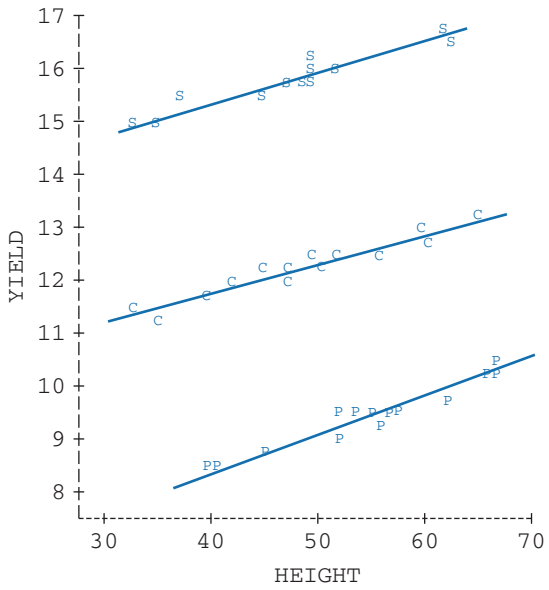


FIGURE 2.4
Plot of plant height versus yield: S = supplemental lighting, C = normal lighting, P = partial shading

of differing size and maturity. Consequently, the height of the plant, a measurable characteristic of plant vigor, was determined at the start of the experiment and will serve as a covariate. This will allow the researcher to adjust the yields of the individual soybean plants depending on the initial size of the plant. On each plant, we record two variables, (x, y) where x is the height of the plant at the beginning of the study and y is the yield of soybeans at the conclusion of the study. To determine whether the covariate has an effect on the response variable, we plot the two variables to assess any possible relationship. If no relationship exists, then the covariate need not be used in the analysis. If the two variables are related, then we must use the techniques of **analysis of covariance** to properly adjust the response variable prior to comparing the mean yields of the three treatments. An initial assessment of the viability of the relationship is simply to plot the response variable versus the covariate with a separate plotting characteristic for each treatment. Figure 2.4 contains this plot for the soybean data.

analysis of covariance

From Figure 2.4, we observe that there appears to be an increasing relationship between the covariate—initial plant height—and the response variable—yield. Also, the three treatments appear to have differing yields; some of the variation in the response variable is related to the initial height as well as to the difference in the amount of lighting the plant received. Thus, we must identify the amount of variation associated with initial height prior to testing for differences in the average yields of the three treatments. We can accomplish this using the techniques of analysis of variance. The analysis of covariance procedures will be discussed in detail in Chapter 16. ■

2.6 RESEARCH STUDY: Exit Polls Versus Election Results

In the beginning of this chapter, we discussed the apparent "discrepancy" between exit polls and the actual voter count during the 2004 presidential election. We will now attempt to answer the following question.

Why were there discrepancies between the exit polls and the election results obtained for the 11 "crucial" states? We will not be able to answer this question definitely, but we can look at some of the issues that pollsters must address when relying on exit polls to accurately predict election results.

First, we need to understand how an exit poll is conducted. We will examine the process as implemented by one such polling company, ***Edison Media Research and Mitofsky International***, as reported on its website. The company conducted exit polls in each state. The state exit poll was conducted at a random sample of polling places among Election Day voters. The polling places are a stratified probability sample of a state. Within each polling place, an interviewer approached every *n*th voter as he or she exited the polling place. Approximately 100 voters completed a questionnaire at each polling place. The exact number depends on voter turnout and the willingness of selected voters to cooperate.

In addition, absentee and/or early voters were interviewed in pre-election telephone polls in a number of states. All samples were random-digit dialing (RDD) selections except for Oregon, which used both RDD and some follow-up calling. Absentee or early voters were asked the same questions as voters at the polling place on Election Day. Results from the phone poll were combined with results from voters interviewed at the polling places. The combination reflects approximately the correct proportion of absentee/early voters and Election Day voters.

The first step in addressing the discrepancies between the exit poll results and actual election tabulation numbers would be to examine the results for all states, not just those thought to be crucial in determining the outcome of the election. These data are not readily available. Next, we would have to make certain that voter fraud was not the cause for the discrepancies. That is the job of the state voter commissions. What can go wrong with exit polls? A number of possibilities exist, including the following:

1. **Nonresponse:** How are the results adjusted for sampled voters refusing to complete the survey? How are the RDD results adjusted for those screening their calls and refusing to participate?
2. **Wording of the questions on the survey:** How were the questions asked? Were they worded in an unbiased, neutral way without leading questions?
3. **Timing of the exit poll:** Were the polls conducted throughout the day at each polling station or just during one time frame?
4. **Interviewer bias:** Were the interviewers unbiased in the way they approached sampled voters?
5. **Influence of election officials:** Did the election officials evenly enforce election laws at the polling booths? Did the officials have an impact on the exit pollsters?
6. **Voter validity:** Did those voters who agreed to be polled give accurate answers to the questions asked?
7. **Agreement with similar pre-election surveys:** Finally, when the exit polls were obtained, did they agree with the most recent pre-election surveys? If not, why not?

Raising these issues is not meant to say that exit polls cannot be of use in predicting actual election results, but they should be used with discretion and with safeguards to mitigate the issues we have addressed as well as other potential problems. But, in the end, it is absolutely essential that no exit poll results be made public until the polls across the country are closed. Otherwise, there is a significant, serious chance

that potential voters may be influenced by the results, thus affecting their vote or, worse, causing them to decide not to vote based on the conclusions derived from the exit polls.

2.7 Summary

The first step in Learning from Data involves defining the problem. This was discussed in Chapter 1. Next, we discussed intelligent data gathering, which involves specifying the objectives of the data-gathering exercise, identifying the variables of interest, and choosing an appropriate design for the survey or experimental study. In this chapter, we discussed various survey designs and experimental designs for scientific studies. Armed with a basic understanding of some design considerations for conducting surveys or scientific studies, you can address how to collect data on the variables of interest in order to address the stated objectives of the data-gathering exercise.

We also drew a distinction between observational and experimental studies in terms of the inferences (conclusions) that can be drawn from the sample data. Differences found between treatment groups from an observational study are said to be *associated with* the use of the treatments; on the other hand, differences found between treatments in a scientific study are said to be *due to* the treatments. In the next chapter, we will examine the methods for summarizing the data we collect.

2.8 Exercises

2.2 Observational Studies

2.1 In the following descriptions of a study, confounding is present. Describe the explanatory and confounding variable in the study and how the confounding may invalidate the conclusions of the study. Furthermore, suggest how you would change the study to eliminate the effect of the confounding variable.

a. A prospective study is conducted to study the relationship between incidence of lung cancer and level of alcohol drinking. The drinking status of 5,000 subjects is determined, and the health of the subjects is then followed for 10 years. The results are given below.

	Lung Cancer		
Drinking Status	**Yes**	**No**	**Total**
Heavy drinker	50	2,150	2,200
Light drinker	30	2,770	2,800
Total	80	4,920	5,000

b. A study was conducted to examine the possible relationship between coronary disease and obesity. The study found that the proportion of obese persons having developed coronary disease was much higher than the proportion of nonobese persons. A medical researcher states that the population of obese persons generally has higher incidences of hypertension and diabetes than the population of nonobese persons.

2.2 In the following descriptions of a study, confounding is present. Describe the explanatory and confounding variable in the study and how the confounding may invalidate the conclusions of the study. Furthermore, suggest how you would change the study to eliminate the effect of the confounding variable.

a. A hospital introduces a new screening procedure to identify patients suffering from a stroke so that a new blood clot medication can be given to the patient during the crucial period of 12 hours after stroke begins. The procedure appears to be very successful because in the first year of its implementation there is a higher rate of total recovery by the patients in comparison to the rate in the previous year for patients admitted to the hospital.

b. A high school mathematics teacher is convinced that a new software program will improve math scores for students taking the SAT. As a method of evaluating her theory, she offers the students an opportunity to use the software on the school's computers during a 1-hour period after school. The teacher concludes the software is effective because the students using the software had significantly higher scores on the SAT than did the students who did not use the software.

2.3 A news report states that minority children who take advanced mathematics courses in high school have a first-year GPA in college that is equivalent to that of white students. The newspaper columnist suggested that the lack of advanced mathematics courses in high school curriculums in inner-city schools was a major cause of the low college success rate of students from inner-city schools. What confounding variables may be present that invalidate the columnist's conclusion?

2.4 A study was conducted to determine if the inclusion of a foreign language requirement in high schools may have a positive effect on students' performance on standardized English exams. From a sample of 100 high schools, 50 of which had a foreign language requirement and 50 of which did not, it was found that the average score on the English proficiency exam was 25% higher for the students having a foreign language requirement. What confounding variables may be present that would invalidate the conclusion that requiring a foreign language in high school increases English language proficiency?

2.3 Sampling Designs for Surveys

Gov. **2.5** The board of directors of a city-owned electric power plant in a large urban city wants to assess the increase in electricity demands due to sources such as hybrid cars, big-screen TVs, and other entertainment devices in the home. There are a number of different sampling plans that can be implemented to survey the residents of the city. What are the relative merits of the following sampling units: individual families, dwelling units (single-family homes, apartment buildings, etc.), and city blocks?

H.R. **2.6** A large auto parts supplier with distribution centers throughout the United States wants to survey its employees concerning health insurance coverage. Employee insurance plans vary greatly from state to state. The company wants to obtain an estimate of the annual health insurance deductible its employees would find acceptable. What sampling plan would you suggest to the company to achieve its goal?

Pol. Sci. **2.7** The circuit judges in a rural county are considering a change in how jury pools are selected for felony trials. They ask the administrator of the courts to assess the county residents' reaction to changing the requirement for membership in the jury pool from the current requirement of all registered voters to a new requirement of all registered voters plus all residents with a current driver's license. The administrator sends questionnaires to a random sample of 1,000 people from the list of registered voters in the county and receives responses from 253 people.

a. What is the population of interest?

b. What is the sampling frame?

c. What possible biases could be present in using the information from the survey?

Psy. **2.8** An evaluation of whether people are truthful in their responses to survey questions was conducted in the following manner. In the first survey, 1,000 randomly selected persons were told during a home visit that the survey was being done to obtain information that would help protect

the drinking water supply in their city. After the short introduction, they were asked if they used a brand of detergent that was biodegradable. In the second survey, 1,000 randomly selected persons were also given the information about safe drinking water during a home visit and then were asked if they used a biodegradable detergent. If they said yes, the interviewer asked to see the box of detergent.

a. What differences do you think will be found in the two estimates of the percentage of households using biodegradable detergents?
b. What types of biases may be introduced into the two types of surveys?

Edu. **2.9** ***Time*** magazine, in an article in the late 1950s, stated that "the average Yaleman, class of 1924, makes $25,111 a year," which, in today's dollars, would be over $150,000. *Time*'s estimate was based on replies to a sample survey questionnaire mailed to those members of the Yale class of 1924 whose addresses were on file with the Yale administration in the late 1950s.

a. What is the survey's population of interest?
b. Were the techniques used in selecting the sample likely to produce a sample that was representative of the population of interest?
c. What are the possible sources of bias in the procedures used to obtain the sample?
d. Based on the sources of bias, do you believe that *Time*'s estimate of the salary of a 1924 Yale graduate in the late 1950s is too high, too low, or nearly the correct value?

2.10 The New York City school district is planning a survey of 1,000 of its 250,000 parents or guardians who have students currently enrolled. They want to assess the parents' opinion about mandatory drug testing of all students participating in any extracurricular activities, not just sports. An alphabetical listing of all parents or guardians is available for selecting the sample. In each of the following descriptions of the method of selecting the 1,000 participants in the survey, identify the type of sampling method used (simple random sampling, stratified sampling, or cluster sampling).

a. Each name is randomly assigned a number. The names with numbers 1 through 1,000 are selected for the survey.
b. The schools are divided into five groups according to grade level taught at the school: K–2, 3–5, 6–7, 8–9, 10–12. Five separate sampling frames are constructed, one for each group. A simple random sample of 200 parents or guardians is selected from each group.
c. The school district is also concerned that the parent's or guardian's opinion may differ depending on the age and sex of the student. Each name is randomly assigned a number. The names with numbers 1 through 1,000 are selected for the survey. The parent is asked to fill out a separate survey for each of their currently enrolled children.

2.11 A professional society, with a membership of 45,000, is designing a study to evaluate its members' satisfaction with the type of sessions presented at the society's annual meeting. In each of the following descriptions of the method of selecting participants in the survey, identify the type of sampling method used (simple random sampling, stratified sampling, or cluster sampling).

a. The society has an alphabetical listing of all its members. It assigns a number to each name and then using a computer software program generates 1,250 numbers between 1 and 45,000. It selects these 1,250 members for the survey.
b. The society is interested in regional differences in its members' opinions. Therefore, it divides the United States into nine regions with approximately 5,000 members per region. It then randomly selects 450 members from each region for inclusion in the survey.
c. The society is composed of doctors, nurses, and therapists, all working in hospitals. There are a total of 450 distinct hospitals. The society decides to conduct onsite in-person interviews, so it randomly selects 20 hospitals and interviews all members working at the selected hospital.

2.12 For each of the following situations, decide what sampling method you would use. Provide an explanation of why you selected a particular method of sampling.

a. A large automotive company wants to upgrade the software on its notebook computers. A survey of 1,500 employees will request information concerning frequently used software applications such as spreadsheets, word processing, e-mail, Internet access, statistical data processing, and so on. A list of employees with their job categories is available.

b. A hospital is interested in what types of patients make use of their emergency room facilities. It is decided to sample 10% of all patients arriving at the emergency room for the next month and record their demographic information along with type of service required, the amount of time the patient waits prior to examination, and the amount of time needed for the doctor to assess the patient's problem.

2.13 For each of the following situations, decide what sampling method you would use. Provide an explanation of why you selected a particular method of sampling.

a. The major state university in the state is attempting to lobby the state legislature for a bill that would allow the university to charge a higher tuition rate than the other universities in the state. To provide a justification, the university plans to conduct a mail survey of its alumni to collect information concerning their current employment status. The university grants a wide variety of different degrees and wants to make sure that information is obtained about graduates from each of the degree types. A 5% sample of alumni is considered sufficient.

b. The Environmental Protection Agency (EPA) is required to inspect landfills in the United States for the presence of certain types of toxic material. The materials were sealed in containers and placed in the landfills. The exact location of the containers is no longer known. The EPA wants to inspect a sample of 100 containers from the 4,000 containers known to be in the landfills to determine if leakage from the containers has occurred.

2.5 Designs for Experimental Studies

Engin.

2.14 The process engineer designed a study to evaluate the quality of plastic irrigation pipes. The study involved a total of 48 pipes; 24 pipes were randomly selected from each of the company's two manufacturing plants. The pipes were heat-treated at one one of four temperatures (175, 200, 225, 250°F). The pipes were chemically treated with one of three types of hardeners (H_1, H_2, H_3). The deviations from the nominal compressive strength were measured at five locations on each of the pipes.

Pipe No.	Plant	Temperature (°F)	Hardener	Pipe No.	Plant	Temperature (°F)	Hardener
1	1	200	H_1	15	2	200	H_3
2	1	175	H_2	16	2	175	H_3
3	2	200	H_1	17	1	200	H_2
4	2	175	H_2	18	1	175	H_1
5	1	200	H_1	19	2	200	H_2
6	1	175	H_2	20	2	175	H_1
7	2	200	H_1	21	1	200	H_2
8	2	175	H_2	22	1	175	H_1
9	1	200	H_3	23	2	200	H_2
10	1	175	H_3	24	2	175	H_1
11	2	200	H_3	25	1	250	H_1
12	2	175	H_3	26	1	225	H_2
13	1	200	H_3	27	2	250	H_1
14	1	175	H_3	28	2	225	H_2

29	1	250	H_1	39	2	250	H_3
30	1	225	H_2	40	2	225	H_3
31	2	250	H_1	41	1	250	H_1
32	2	225	H_2	42	1	225	H_2
33	1	250	H_3	43	2	250	H_1
34	1	225	H_3	44	2	225	H_2
35	2	250	H_3	45	1	250	H_1
36	2	225	H_3	46	1	225	H_2
37	1	250	H_3	47	2	250	H_1
38	1	225	H_3	48	2	225	H_2

Identify each of the following components of the experimental design.

a. Factors
b. Factor levels
c. Blocks
d. Experimental unit
e. Measurement unit
f. Replications
g. Covariates
h. Treatments

In the descriptions of experiments given in Exercises 2.15–2.18, identify the important features of each design. Include as many of the components listed in Exercise 2.14 as needed to adequately describe the design.

Ag. **2.15** A horticulturist is measuring the vitamin C concentration in oranges in an orchard on a research farm in south Texas. He is interested in the variation in vitamin C concentration across the orchard, across the productive months, and within each tree. He divides the orchard into eight sections and randomly selects a tree from each section during October–May, the months in which the trees are in production. During each month, he selects from each of the eight trees 10 oranges near the top of the tree, 10 oranges near the middle of the tree, and 10 oranges near the bottom of the tree. The horticulturist wants to monitor the vitamin C concentration across the productive season and determine if there is a substantial difference in vitamin C concentration in oranges at various locations in the tree.

Med. **2.16** A medical study is designed to evaluate a new drug, D_1, for treating a particular illness. There is a widely used treatment, D_2, for this disease to which the new drug will be compared. A placebo will also be included in the study. The researcher has selected 10 hospitals for the study. She does a thorough evaluation of the hospitals and concludes that there may be aspects of the hospitals that may result in the elevation of responses at some of the hospitals. Each hospital has six wards of patients. She will randomly select six patients in each ward to participate in the study. Within each hospital, two wards are randomly assigned to administer D_1, two wards to administer D_2, and two wards administer the placebo. All six patients in each of the wards will be given the same treatment. Age, BMI, blood pressure, and a measure of degree of illness are recorded for each patient upon entry into the hospital. The response is an assessment of the degree of illness after 6 days of treatment.

Med. **2.17** In place of the design described in Exercise 2.16, make the following change. Within each hospital, the three treatments will be randomly assigned to the patients, with two patients in each ward receiving D_1, two patients receiving D_2, and two patients receiving the placebo.

Edu. **2.18** Researchers in an education department at a large state university have designed a study to compare the math abilities of students in junior high. They will also examine the impact of three types of schools—public, private nonparochial, and parochial—on the scores the students receive in a standardized math test. Two large cities in each of four geographical regions of the United States were selected for the study. In each city, one school of each of the three types was randomly selected, and a single eighth-grade class was randomly selected within each school.

The scores on the test were recorded for each student in the selected classrooms. The researcher was concerned about differences in socio-economie status among the 8 cities, so she obtained a measure of socioeconomic status for each of the students that participated in the study.

Bio. **2.19** A research specialist for a large seafood company plans to investigate bacterial growth on oysters and mussels subjected to three different storage temperatures. Nine cold-storage units are available. She plans to use three storage units for each of the three temperatures. One package of oysters and one package of mussels will be stored in each of the storage units for 2 weeks. At the end of the storage period, the packages will be removed and the bacterial count made for two samples from each package. The treatment factors of interest are temperature (levels: 0, 5, 10°C) and seafood (levels: oysters, mussels). She will also record the bacterial count for each package prior to placing seafood in the cooler. Identify each of the following components of the experimental design.

- **a.** Factors
- **b.** Factor levels
- **c.** Blocks
- **d.** Experimental unit
- **e.** Measurement unit
- **f.** Replications
- **g.** Treatments

In Exercises 2.20–2.22, identify whether the design is a completely randomized design, randomized complete block design, or Latin square design. If there is a factorial structure for the treatments, specify whether it has a two-factor or three-factor structure. If the measurement units are different from the experimental units, identify both.

Ag. **2.20** The researchers design an experiment to evaluate the effect of applying fertilizer having varying levels of nitrogen, potassium, and phosphorus on the yields of orange trees. There were three, four, and three different levels of N, P, and K, respectively, yielding 36 distinct combinations. Ten orange groves were randomly selected for the experiment. Each grove was then divided into 36 distinct plots, and the 36 fertilizer combinations were randomly assigned to the plots within each grove. The yield of five randomly selected trees in each plot is recorded to assess the variation within each of the 360 plots.

Bus. **2.21** A company is planning on purchasing a software program to manage its inventory. Five vendors submit bids on supplying the inventory control software. In order to evaluate the effectiveness of the software, the company's personnel decide to evaluate the software by running each of the five software packages at each of the company's 10 warehouses. The number of errors produced by each of the software packages is recorded at each of the warehouses.

Sci. **2.22** Four different glazes are applied at two different thicknesses to clay pots. The kiln used in the glazing can hold eight pots at a time, and it takes 1 day to apply the glazes. The experimenter wanted eight replications of the experiment. Since conditions in the kiln vary somewhat from day to day, the experiment was conducted over an 8-day period. The experiment is conducted so that each combination of a thickness and type of glaze is randomly assigned to one pot in the kiln each day.

Bus. **2.23** A bakery wants to evaluate new recipes for carrot cake. It decides to ask a random sample of regular customers to evaluate the recipes by tasting samples of the cakes. After a customer tastes a sample of the cake, the customer will provide scores for several characteristics of the cake, and these scores are then combined into a single overall score for the recipe. Thus, from each customer, a single numeric score is recorded for each recipe. The taste-testing literature indicates that in this type of study some consumers tend to give all samples low scores and others tend to give all samples high scores.

- **a.** There are two possible experimental designs. Design A would use a random sample of 100 customers. From this group, 20 would be randomly assigned to each of the five recipes, so that each customer tastes only one recipe. Design B would use a random sample of 100 customers with each customer tasting all five recipes, the recipes being presented in a random order for each customer. Which design would you recommend? Justify your answer.

b. The manager of the bakery asked for a progress report on the experiment. The person conducting the experiment replied that one recipe tasted so bad that she eliminated it from the analysis. Is this a problem for the analysis if Design B was used? Why or why not? Would it have been a problem if Design A was used? Why or why not?

Supplementary Exercises

H.R. **2.24** A large healthcare corporation is interested in the number of employees who devote a substantial amount of time to providing care for elderly relatives. The corporation wants to develop a policy with respect to the number of sick days an employee can use to provide care to elderly relatives. The corporation has thousands of employees, so it decides to have a sample of employees fill out a questionnaire.

a. How would you define *employee?* Should only full-time workers be considered?
b. How would you select the sample of employees?
c. What information should be collected from the workers?

Bus. **2.25** The school of nursing at a university is developing a long-term plan to determine the number of faculty members that may be needed in future years. Thus, it needs to determine the future demand for nurses in the areas in which many of the graduates find employment. The school decides to survey medical facilities and private doctors to assist in determining the future nursing demand.

a. How would you obtain a list of private doctors and medical facilities so that a sample of doctors could be selected to fill out a questionnaire?
b. What are some of the questions that should be included on the questionnaire?
c. How would you determine the number of nurses who are licensed but not currently employed?
d. What are some possible sources for determining the population growth and health risk factors for the areas in which many of the nurses find employment?
e. How could you sample the population of healthcare facilities and types of private doctors so as not to exclude any medical specialties from the survey?

2.26 Consider the yields given in Table 2.7. In this situation, there is no interaction. Show that the one-at-a-time approach would result in the experimenter finding the best combination of nitrogen and phosphorus—that is, the combination producing maximum yield. Your solution should include the five combinations you would use in the experiment.

2.27 The population values that would result from running a 2×3 factorial treatment structure are given in the following table. Note that two values are missing. If there is *no interaction* between the two factors, determine the missing values.

	Factor 2		
Factor 1	**I**	**II**	**III**
A	25	45	
B		30	50

Vet. **2.28** An experiment is designed to evaluate the effect of different levels of exercise on the health of dogs. The two levels are L_1—1-mile walk every day and L_2—2-mile walk every other day. At the end of a 3-month study period, each dog will undergo measurements of respiratory and cardiovascular fitness from which a fitness index will be computed. There are 16 dogs available for the study. They are all in good health and are of the same general size, which is within the normal range for their breed. The following table provides information about the sex and age of the 16 dogs.

Dog	Sex	Age	Dog	Sex	Age
1	F	5	9	F	8
2	F	3	10	F	9
3	M	4	11	F	6
4	M	7	12	M	8
5	M	2	13	F	2
6	M	3	14	F	1
7	F	5	15	M	6
8	M	9	16	M	3

a. How would you group the dogs prior to assigning the treatments to obtain a study having as small an experimental error as possible? List the dogs in each of your groups.

b. Describe your procedure for assigning the treatments to the individual dogs using a random number generator.

Bus. **2.29** Four cake recipes are to be compared for moistness. The researcher will conduct the experiment by preparing and then baking the cake. Each preparation of a recipe makes only one cake. All recipes require the same cooking temperature and the same length of cooking time. The oven is large enough that four cakes may be baked during any one baking period, in positions P_1 through P_4, as shown here.

P_1	P_2
P_3	P_4

a. Discuss an appropriate experimental design and randomization procedure if there are to be r cakes for each recipe.

b. Suppose the experimenter is concerned that significant differences could exist due to the four baking positions in the oven (front vs. back, left side vs. right side). Is your design still appropriate? If not, describe an appropriate design.

c. For the design or designs described in (b), suggest modifications if there are five recipes to be tested but only four cakes may be cooked at any one time.

Bio. **2.30** A forester wants to estimate the total number of trees on a tree farm that have a diameter exceeding 12 inches. Because the farm contains too many trees to facilitate measuring all of them, she uses Google Earth to divide the farm into 250 rectangular plots of approximately the same area. An examination of the plots reveals that 27 of the plots have a sizable portion of their land under water. The forester excluded the 27 "watery" plots for the study. She then randomly selected 42 of the remaining 223 plots and counted all the trees having a diameter exceeding 12 inches on the 42 selected plots.

a. What is the sampling frame for this study?

b. How does the sampling frame differ from the population of interest, if at all?

c. What biases may exist in the estimate of the number of trees having a diameter greater than 12 inches based on the collected data?

Engin. **2.31** A transportation researcher is funded to estimate the proportion of automobile tires with an unsafe tread thickness in a small northern state. The researcher randomly selects one month during each of the four seasons for taking the measurements. During each of the four selected months, the researcher randomly selects 500 cars from the list of registered cars in the state and then measures the tread thickness of the four tires on each of the selected cars.

a. What is the population of interest?

b. What is the sampling frame?

c. What biases if any may result from using the data from this study to obtain the estimated proportion of cars with an unsafe thread thickness?

Gov. **2.32** The department of agriculture in a midwestern state wants to estimate the amount of corn produced in the state that is used to make ethanol. There are 50,000 farms in the state that produce corn. The farms are classified into four groups depending on the total number of acres planted in corn. A random sample of 500 farms is selected from each of the four groups, and the amount corn used to generate ethanol is determined for each of the 2,000 selected farms.

- **a.** What is the population of interest?
- **b.** What is the sampling frame?
- **c.** What type of sampling plan is being used in this study?
- **d.** What biases if any may result from using the data from this study to obtain an estimate of the amount of corn used to produce ethanol?

2.33 Discuss the relative merits of using personal interviews, telephone interviews, and mailed questionnaires as data collection methods for each of the following situations:

- **a.** A television executive wants to estimate the proportion of viewers in the country who are watching the network at a certain hour.
- **b.** A newspaper editor wants to survey the attitudes of the public toward the type of news coverage offered by the paper.
- **c.** A city commissioner is interested in determining how homeowners feel about a proposed zoning change.
- **d.** A county health department wants to estimate the proportion of dogs that have had rabies shots within the last year.

Soc. **2.34** A ***Yankelovich, Skelly, and White*** poll taken in the fall of 1984 showed that one-fifth of the 2,207 people surveyed admitted to having cheated on their federal income taxes. Do you think that this fraction is close to the actual proportion who cheated? Why? (Discuss the difficulties of obtaining accurate information on a question of this type.)

PART 3

Summarizing Data

CHAPTER 3

Data Description

3.1 Introduction and Abstract of Research Study

In the previous chapter, we discussed how to gather data intelligently for an experiment or survey, Step 2 in Learning from Data. We turn now to Step 3, summarizing the data.

The field of statistics can be divided into two major branches: descriptive statistics and inferential statistics. In both branches, we work with a set of measurements. For situations in which data description is our major objective, the set of measurements available to us is frequently the entire population. For example, suppose that we wish to describe the distribution of annual incomes for all families registered in the 2000 census. Because all these data are recorded and are available on computer tapes, we do not need to obtain a random sample from the population; the complete set of measurements is at our disposal. Our major problem is in organizing, summarizing, and describing these data—that is, making sense of the data. Similarly, vast amounts of monthly, quarterly, and yearly data of medical costs are available for the managed healthcare industry, HMOs.

These data are broken down by type of illness, age of patient, inpatient or outpatient care, prescription costs, and out-of-region reimbursements, along with many other types of expenses. However, in order to present such data in formats useful to HMO managers, congressional staffs, doctors, and the consuming public, it is necessary to organize, summarize, and describe the data. Good descriptive statistics enable us to make sense of the data by reducing a large set of measurements to a few summary measures that provide a good, rough picture of the original measurements.

In situations in which we are unable to observe all units in the population, a sample is selected from the population, and the appropriate measurements are made. We use the information in the sample to draw conclusions about the population from which the sample was drawn. However, in order for these inferences about the population to have a valid interpretation, the sample should be a random sample of one of the forms discussed in Chapter 2. During the process of making inferences, we also need to organize, summarize, and describe the data.

Following the tragedies that occurred on September 11, 2001, the Transportation Security Administration (TSA) was created to strengthen the security of the nation's transportation systems. TSA has the responsibility to secure the nation's airports and screens all commercial airline passengers and baggage. Approximately 1.8 million passengers pass through our nation's airports every day. TSA attempts to provide the highest level of security and customer service to all who pass through our screening checkpoints. However, if every passenger was physically inspected by a TSA officer, the delay in the airports would be unacceptable to the traveling public. Thus, TSA focuses its resources at security checkpoints by applying new intelligence-driven, risk-based screening procedures and enhancing its use of technology. Instead of inspecting every passenger, TSA employs a system of randomly selecting passengers for screening together with random and unpredictable security measures throughout the airport. No individual will be guaranteed expedited screening in order to retain a certain element of randomness to prevent terrorists from gaming the system.

Similarly, in order to monitor changes in the purchasing power of consumers' income, the federal government uses the Consumer Price Index (CPI) to measure the average change in prices over time in a market basket of goods and services purchased by urban wage earners. The current CPI is based on prices of food, clothing, shelter, fuels, transportation fares, charges for doctors' and dentists' services, drugs, and so on, purchased for day-to-day living. Each month the Bureau of Labor Statistics (BLS) scientifically samples approximately 80,000 goods and services purchased by consumers. The CPI is estimated from these samples of consumer purchases; it is not a complete measure of price change. Consequently, the index results may deviate slightly from those that would be obtained if all consumer transactions were recorded. This is called sampling error. These estimation or sampling errors are statistical limitations of the index. A different kind of error in the CPI can occur when, for example, a respondent provides BLS field representatives with inaccurate or incomplete information. This is called nonsampling error.

A third situation involves an experiment in which a drug company wants to study the effects of two factors on the level of blood sugar in diabetic patients. The factors are the type of drug (a new drug and two drugs currently being used) and the method of administering the drug to the diabetic patient (two different delivery modes). The experiment involves randomly selecting a method of administering the drug and randomly selecting a type of drug and then giving the drug to the patient. The fasting blood sugar of the patient is then recorded at the time the

patient receives the drug and at 6-hour intervals over a 2-day period of time. The six unique combinations of type of drug and method of delivery are given to 10 different patients. In this experiment, the drug company wants to make inferences from the results of the experiment to determine if the new drug is commercially viable. In many experiments of this type, the use of proper graphical displays provides valuable insights to the scientists with respect to identifying unusual occurrences and making comparisons of the responses to the different treatment combinations.

Whether we are describing an observed population or using sampled data to draw an inference from the sample to the population, an insightful description of the data is an important step in drawing conclusions from it. No matter what our objective, statistical inference or population description, we must first adequately describe the set of measurements at our disposal.

The two major methods for describing a set of measurements are graphical techniques and numerical descriptive techniques. Section 3.3 deals with graphical methods for describing data on a single variable. In Sections 3.4, 3.5, and 3.6, we discuss numerical techniques for describing data. The final topics on data description are presented in Section 3.7, in which we consider a few techniques for describing (summarizing) data on more than one variable. A research study involving the evaluation of primary school teachers will be used to illustrate many of the summary statistics and graphs introduced in this chapter.

Abstract of Research Study: Controlling for Student Background in the Assessment of Teaching

By way of background, there was a movement to introduce achievement standards and school/teacher accountability in the public schools of our nation long before the No Child Left Behind bill was passed by the Congress during the first term of President George W. Bush. However, even after an important federal study entitled *A Nation at Risk (National Commission on Excellence in Education*, 1983) spelled out the grave trend toward mediocrity in our schools and the risk this poses for the future, Presidents Ronald Reagan, George H. W. Bush, and Bill Clinton did not venture into this potentially sensitive area to champion meaningful change.

Many politicians, teachers, and educational organizations have criticized the No Child Left Behind (NCLB) legislation, which requires rigid testing standards in exchange for money to support low-income students. A recent survey conducted by the Educational Testing Service (ETS) with bipartisan sponsorship from the Congress showed the following:

- Those surveyed identified the value of our education as the most important source of the United States' success in the world. (Also included on the list of alternatives were our military strength, our geographical and natural resources, our democratic system of government, our entrepreneurial spirit, etc.)
- 45% of the parents surveyed viewed the NCLB reforms favorably; 34% viewed them unfavorably.
- Only 19% of the high school teachers surveyed viewed the NCLB reforms favorably, while 75% viewed them unfavorably.

Given the importance placed on education, the difference or gap between the responses of parents and those of educators is troubling. The tone of much of the criticism seems to run against the empirical results seen to date with the NCLB program. For example, in 2004 the Center on Education Policy, an independent

research organization, reported that 36 of 49 (73.5%) schools surveyed showed improvement in student achievement.

One of the possible sources of criticism coming from the educators is that there is a risk of being placed on a "watch list" if the school does not meet the performance standards set. This would reflect badly on the teachers, the school, and the community. But another important source of the criticism voiced by the teachers and reflected in the gap between what parents and teachers favor relates to the performance standards themselves. In the previously mentioned ETS survey, those polled were asked whether the same standard should be used for all students of a given grade, regardless of their background, because of the view that it is wrong to have lower expectations for students from disadvantaged backgrounds. The opposing view is that it is not reasonable to expect teachers to be able to bring the achievement for disadvantaged students to the same level as that of students from more affluent areas. While more than 50% of the parents favored a single standard, only 25% of the teachers suggested this view.

Next, we will examine some data that may offer a way to improve the NCLB program while maintaining the important concepts of performance standards and accountability.

In an article in the Spring 2004 issue of the **Journal of Educational and Behavioral Statistics, *"An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance,"*** by Tekwe et al., data were presented from three elementary school grade cohorts (third–fifth grades) in 1999 in a medium-sized Florida school district with 22 elementary schools. The data are given in Table 3.1. The minority

TABLE 3.1
Assessment of elementary school performance

	Third Grade				
School	**Math**	**Reading**	**%Minority**	**%Poverty**	***N***
1	166.4	165.0	79.2	91.7	48
2	159.6	157.2	73.8	90.2	61
3	159.1	164.4	75.4	86.0	57
4	155.5	162.4	87.4	83.9	87
5	164.3	162.5	37.3	80.4	51
6	169.8	164.9	76.5	76.5	68
7	155.7	162.0	68.0	76.0	75
8	165.2	165.0	53.7	75.8	95
9	175.4	173.7	31.3	75.6	45
10	178.1	171.0	13.9	75.0	36
11	167.1	169.4	36.7	74.7	79
12	177.1	172.9	26.5	63.2	68
13	174.2	172.7	28.3	52.9	191
14	175.6	174.9	23.7	48.5	97
15	170.8	174.9	14.5	39.1	110
16	175.1	170.1	25.6	38.4	86
17	182.8	181.4	22.9	34.3	70
18	180.3	180.6	15.8	30.3	165
19	178.8	178.0	14.6	30.3	89
20	181.4	175.9	28.6	29.6	98
21	182.8	181.6	21.4	26.5	98
22	186.1	183.8	12.3	13.8	130

(continued)

TABLE 3.1
Assessment of elementary school performance *(continued)*

Fourth Grade

School	Math	Reading	%Minority	%Poverty	*N*
1	181.1	177.0	78.9	89.5	38
2	181.1	173.8	75.9	79.6	54
3	180.9	175.5	64.1	71.9	64
4	169.9	166.9	94.4	91.7	72
5	183.6	178.7	38.6	61.4	57
6	178.6	170.3	67.9	83.9	56
7	182.7	178.8	65.8	63.3	79
8	186.1	180.9	48.0	64.7	102
9	187.2	187.3	33.3	62.7	51
10	194.5	188.9	11.1	77.8	36
11	180.3	181.7	47.4	70.5	78
12	187.6	186.3	19.4	59.7	72
13	194.0	189.8	21.6	46.2	171
14	193.1	189.4	28.8	36.9	111
15	195.5	188.0	20.2	38.3	94
16	191.3	186.6	39.7	47.4	78
17	200.1	199.7	23.9	23.9	67
18	196.5	193.5	22.4	32.8	116
19	203.5	204.7	16.0	11.7	94
20	199.6	195.9	31.1	33.3	90
21	203.3	194.9	23.3	25.9	116
22	206.9	202.5	13.1	14.8	122

Fifth Grade

School	Math	Reading	%Minority	%Poverty	*N*
1	197.1	186.6	81.0	92.9	42
2	194.9	200.1	83.3	88.1	42
3	192.9	194.5	56.0	80.0	50
4	193.3	189.9	92.6	75.9	54
5	197.7	199.6	21.7	67.4	46
6	193.2	193.6	70.4	76.1	71
7	198.0	200.9	64.1	67.9	78
8	205.2	203.5	45.5	61.0	77
9	210.2	223.3	34.7	73.5	49
10	204.8	199.0	29.4	55.9	34
11	205.7	202.8	42.3	71.2	52
12	201.2	207.8	15.8	51.3	76
13	205.2	203.3	19.8	41.2	131
14	212.7	211.4	26.7	41.6	101
15	—	—	—	—	—
16	209.6	206.5	22.4	37.3	67
17	223.5	217.7	14.3	30.2	63
18	222.8	218.0	16.8	24.8	137
19	—	—	—	—	—
20	228.1	222.4	20.6	23.5	102
21	221.0	221.0	10.5	13.2	114
22	—	—	—	—	—

Source: *Tekwe, C., R. Carter, C. Ma, J. Algina, M. Lucas, J. Roth, M. Ariet, T. Fisher, and M. Resnick. (2004), "An empirical comparison of statistical models for value-added assessment of school performance." Journal of Educational and Behavioral Statistics 29, 11–36.*

status of a student was defined as black or non-black race. In this school district, almost all students are non-Hispanic blacks or whites. Most of the relatively small numbers of Hispanic students are white. Most students of other races are Asian but are relatively few in number. They were grouped in the minority category because of the similarity of their test score profiles. Poverty status was based on whether or not the student received a free or reduced lunch subsidy. The math and reading scores are from the Iowa Test of Basic Skills. The number of students by class in each school is given by N in the table.

The superintendent of the schools presented the school board members with the data, and they wanted an assessment of whether poverty and minority status had any effect on the math and reading scores. Just looking at the data in the table presented very little insight to answering this question. At the end of this chapter, we will present a discussion of what types of graphs and summary statistics would be beneficial to the school board in reaching a conclusion about the impact of these two variables on student performance.

3.2 Calculators, Computers, and Software Systems

Electronic calculators can be great aids in performing some of the calculations mentioned later in this chapter, especially for small data sets. For larger data sets, even hand-held calculators are of little use because of the time required to enter data. A computer can help in these situations. Specific programs or more general software systems can be used to perform statistical analyses almost instantaneously even for very large data sets after the data are entered into the computer. It is not necessary to know computer programming to make use of specific programs or software systems for planned analyses—most provide pull-down menus that lead the user through the analysis of choice.

Many statistical software packages are available. A few of the more commonly used are SAS, SPSS, Minitab, R, JMP, and STATA. Because a software system is a group of programs that work together, it is possible to obtain plots, data descriptions, and complex statistical analyses in a single job. Most people find that they can use any particular system easily, although they may be frustrated by minor errors committed on the first few tries. The ability of such packages to perform complicated analyses on large amounts of data more than repays the initial investment of time and irritation.

In general, to use a system you need to learn about only the programs in which you are interested. Typical steps in a job involve describing your data to the software system, manipulating your data if they are not in the proper format or if you want a subset of your original data set, and then invoking the appropriate set of programs or commands particular to the software system you are using.

Because this isn't a text on computer use, we won't spend additional time and space on the mechanics, which are best learned by doing. Our main interest is in interpreting the output from these programs. The designers of these programs tend to include in the output everything that a user could conceivably want to know; as a result, in any particular situation, some of the output is irrelevant. When reading computer output, look for the values you want; if you don't need or don't understand an output statistic, don't worry. Of course, as you learn more about statistics, more of the output will be meaningful. In the meantime, look for what you need and disregard the rest.

There are dangers in using such packages carelessly. A computer is a mindless beast and will do anything asked of it, no matter how absurd the result might be. For instance, suppose that the data include age, gender (1 = female, 2 = male), political party (1 = Democrat, 2 = Republican, 3 = Green, 4 = Libertarian, 5 = Other, 6 = None), and monthly income of a group of people. If we asked the computer to calculate averages, we would get averages for the variables gender and political party, as well as for age and monthly income, even though these averages are meaningless. For example, suppose a random sample of 100 people identifies their political party as follows: 30 respond Democrat = 1, 30 respond Republican = 2, 10 respond Green = 3, 10 respond Libertarian = 4, 10 respond Other = 5, and 10 respond None = 6. The average of the 100 numbers would be 2.7, which would be a green republican, that is, it would have absolutely no meaning with respect to the "average" political affiliation of the group of 100 people. Used intelligently, these packages are convenient, powerful, and useful—but be sure to examine the output from any computer run to make certain the results make sense. Did anything go wrong? Was something overlooked? In other words, be *skeptical.* One of the important acronyms of computer technology still holds—namely, GIGO: garbage in, garbage out.

Throughout the textbook, we will use computer software systems to do most of the more tedious calculations of statistics *after* we have explained how the calculations can be done. Used in this way, computers (and associated graphical and statistical analysis packages) will enable us to spend additional time on interpreting the results of the analyses rather than on doing the analyses.

3.3 Describing Data on a Single Variable: Graphical Methods

After the measurements of interest have been collected, ideally the data are organized, displayed, and examined by using various graphical techniques. As a general rule, the data should be arranged into categories so that *each measurement is classified into one, and only one, of the categories.* This procedure eliminates any ambiguity that might otherwise arise when categorizing measurements. For example, suppose a sex discrimination lawsuit is filed. The law firm representing the plaintiffs needs to summarize the salaries of all employees in a large corporation. To examine possible inequities in salaries, the law firm decides to summarize the 2005 yearly income rounded to the nearest dollar for all female employees into the categories listed in Table 3.2.

The yearly salary of each female employee falls into one, and only one, income category. However, if the income categories had been defined as shown in

TABLE 3.2
Format for summarizing salary data

Income Level	Salary
1	less than $20,000
2	$20,000 to $39,999
3	$40,000 to $59,999
4	$60,000 to $79,999
5	$80,000 to $99,999
6	$100,000 or more

TABLE 3.3
Format for summarizing salary data

Income Level	Salary
1	less than $20,000
2	$20,000 to $40,000
3	$40,000 to $60,000
4	$60,000 to $80,000
5	$80,000 to $100,000
6	$100,000 or more

Table 3.3, then there would be confusion as to which category should be checked. For example, an employee earning $40,000 could be placed in either category 2 or category 3. To reiterate: If the data are organized into categories, it is important to define the categories so that a measurement can be placed into only one category.

When data are organized according to this general rule, there are several ways to display the data graphically. The first and simplest graphical procedure for data organized in this manner is the **pie chart.** It is used to display the percentage of the total number of measurements falling into each of the categories of the variable by partitioning a circle (similar to slicing a pie).

pie chart

The data of Table 3.4 represent a summary of a study to determine which types of employment may be the most dangerous to their employees. Using data from the National Safety Council, it was reported that in 1999, approximately 3,240,000 workers suffered disabling injuries (an injury that results in death or some degree of physical impairment or that renders the employee unable to perform regular activities for a full day beyond the day of the injury). Each of the 3,240,000 disabled workers was classified according to the industry group in which he or she was employed.

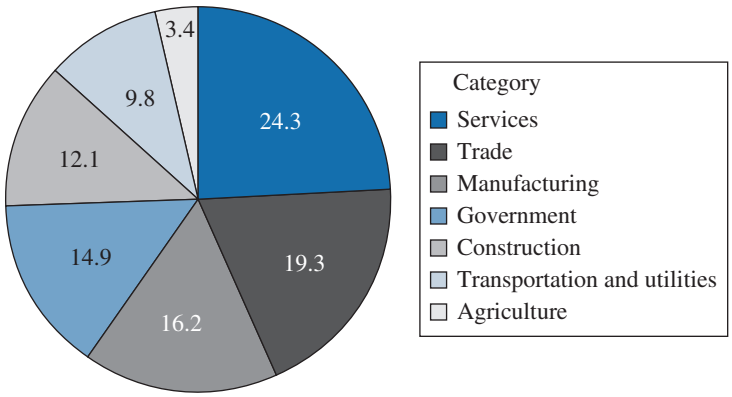
Although you can scan the data in Table 3.4, the results are more easily interpreted by using a pie chart. From Figure 3.1, we can make certain inferences about which industries have the highest number of injured employees and thus may require a closer scrutiny of their practices. For example, the services industry had nearly one-quarter, 24.3%, of all disabling injuries during 1999, whereas government employees constituted only 14.9%. At this point, we must carefully consider what is being displayed in both Table 3.4 and Figure 3.1. They show the numbers of disabling injuries, but these figures do not take into account the numbers of workers employed in the various industry groups. To realistically reflect the risk of a disabling injury to the employees in each of the industry groups, we need to

TABLE 3.4
Disabling injuries by industry group

Industry Group	Number of Disabling Injuries (in 1,000s)	Percent of Total
Agriculture	130	3.4
Construction	470	12.1
Manufacturing	630	16.2
Transportation & utilities	300	9.8
Trade	380	19.3
Services	750	24.3
Government	580	14.9

Source: ***U.S. Census Bureau. (2002), Statistical Abstract of the United States, 122nd ed. Washington, D.C.: U.S. Government Printing Office 2001.***

FIGURE 3.1
Pie chart for the data of Table 3.4



take into account the total number of employees in each of the industries. A rate of disabling injury could then be computed that would be a more informative index of the risk to a worker employed in each of the groups. For example, although the services group had the highest percentage of workers with a disabling injury, it also had the largest number of workers. Taking into account the number of workers employed in each of the industry groups, the services group had the lowest rate of disabling injuries in the seven groups. This illustrates the necessity of carefully examining tables of numbers and graphs prior to drawing conclusions.

Another variation of the pie chart is shown in Figure 3.2. It shows the loss of market share by PepsiCo as a result of the switch by a major fast-food chain from Pepsi to Coca-Cola for its fountain drink sales. In summary, the pie chart can be used to display percentages associated with each category of the variable. The following guidelines should help you to obtain clarity of presentation in pie charts.

Guidelines for Constructing Pie Charts

1. Choose a small number (five or six) of categories for the variable because too many make the pie chart difficult to interpret.
2. Whenever possible, construct the pie chart so that percentages are in either ascending or descending order.

FIGURE 3.2
Estimated U.S. market share before and after switch in soft drink accounts

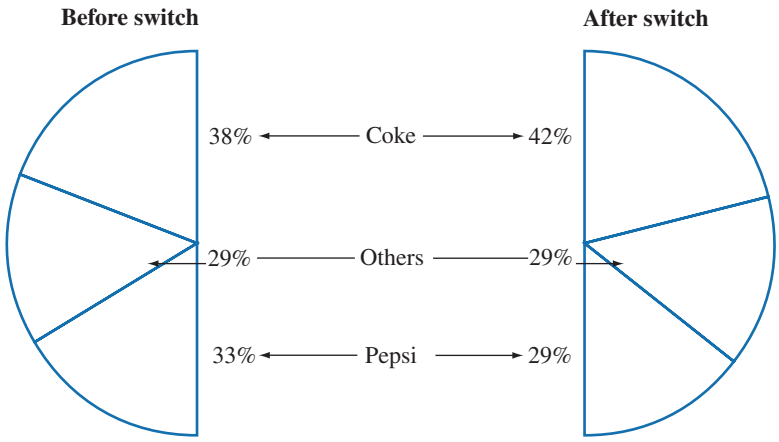


FIGURE 3.3
Number of workers employed by major foreign investors by country

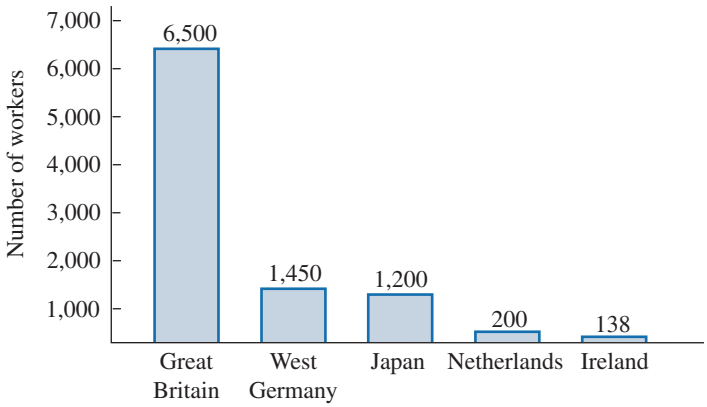
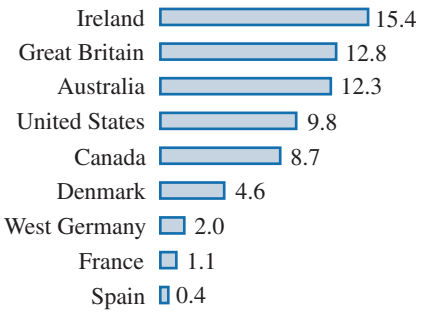
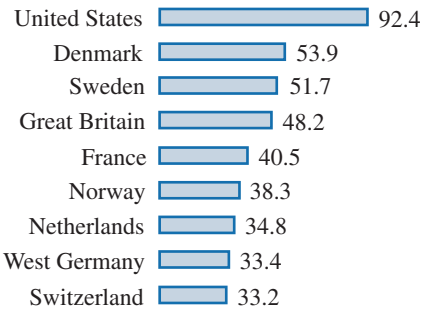


FIGURE 3.4
Greatest per capita consumption by country



(a) Breakfast cereals (in pounds)



(b) Frozen foods (in pounds)

bar chart

A second graphical technique is the **bar chart,** or bar graph. Figure 3.3 displays the number of workers employed in the Cincinnati, Ohio, area by major foreign investors by country. There are many variations of the bar chart. Sometimes the bars are displayed horizontally, as in Figures 3.4(a) and (b). They can also be used to display data across time, as in Figure 3.5. Bar charts are relatively easy to construct if you use the following guidelines.

Guidelines for Constructing Bar Charts

1. Label frequencies on one axis and categories of the variable on the other axis.
2. Construct a rectangle at each category of the variable with a height equal to the frequency (number of observations) in the category.
3. Leave a space between each category to connote distinct, separate categories and to clarify the presentation.

frequency histogram, relative frequency histogram

The next two graphical techniques that we will discuss are the **frequency histogram** and the **relative frequency histogram.** Both of these graphical techniques are applicable only to quantitative (measured) data. As with the pie chart, we must organize the data before constructing a graph.

Gulf Coast ticks are significant pests of grazing cattle that require new strategies of population control. Some particular species of ticks not only are the source

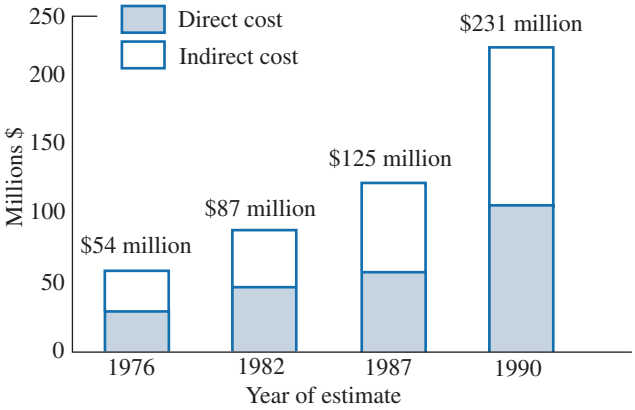


FIGURE 3.5
Estimated direct and indirect costs for developing a new drug by selected years

of considerable economic losses to the cattle industry due to weight loss in the cattle but also are recognized vectors for a number of diseases in cattle. An entomologist carries out an experiment to investigate whether a new repellant for ticks is effective in preventing ticks from attaching to grazing cattle. The researcher determines that 100 cows will provide sufficient information to validate the results of the experiment and convince a commercial enterprise to manufacture and market the repellant. (In Chapter 5, we will present techniques for determining the appropriate sample size for a study to achieve specified goals.) The scientist then exposes the cows to a specified number of ticks in a laboratory setting and records the number of attached ticks after 1 hour of exposure. The average number of attached ticks on cows using a currently marketed repellant is 34 ticks. The scientist wants to demonstrate that using the new repellant will result in a reduction of the average number of attached ticks. The numbers of attached ticks for the 100 cows are presented in Table 3.5.

An initial examination of the tick data reveals that the largest number of ticks is 42 and the smallest is 17. Although we might examine the table very closely to determine whether the number of ticks per cow is substantially less than 34, it is difficult to describe how the measurements are distributed along the interval 17 to 42. One way to facilitate the description is to organize the data in a **frequency table.**

frequency table

To construct a frequency table, we begin by dividing the range from 17 to 42 into an arbitrary number of subintervals called **class intervals.** The number of subintervals chosen depends on the number of measurements in the set, but we generally recommend using from 5 to 20 class intervals. The more data we have, the larger the number of classes we tend to use. The guidelines given here can be used for constructing the appropriate class intervals.

class intervals

TABLE 3.5
Number of attached ticks

17	18	19	20	20	20	21	21	21	22	22	22	22	23	23
23	24	24	24	24	24	25	25	25	25	25	25	25	26	26
27	27	27	27	27	27	28	28	28	28	28	28	28	28	28
28	28	29	29	29	29	29	29	29	29	29	29	30	30	30
30	30	30	30	30	31	31	31	31	31	31	32	32	32	32
32	32	32	32	33	33	33	34	34	34	34	35	35	35	36
36	36	36	37	37	38	39	40	41	42					

Guidelines for Constructing Class Intervals

1. Divide the *range* of the measurements (the difference between the largest and the smallest measurements) by the approximate number of class intervals desired. Generally, we want to have from 5 to 20 class intervals.
2. After dividing the range by the desired number of class intervals, round the resulting number to a convenient (easy to work with) unit. This unit represents a common width for the class intervals.
3. Choose the first class interval so that it contains the smallest measurement. It is also advisable to choose a starting point for the first interval so that no measurement falls on a point of division between two class intervals, which eliminates any ambiguity in placing measurements into the class intervals. (One way to do this is to choose boundaries to one more decimal place than the data.)

For the data in Table 3.5,

$$\text{range} = 42 - 17 = 25$$

Assume that we want to have approximately 10 subintervals. Dividing the range by 10 and rounding to a convenient unit, we have $25/10 = 2.5$. Thus, the class interval width is 2.5.

It is convenient to choose the first interval to be 16.25–18.75, the second to be 18.75–21.25, and so on. Note that the smallest measurement, 17, falls in the first interval and that no measurement falls on the endpoint of a class interval. (See Tables 3.5 and 3.6.)

Having determined the class interval, we construct a frequency table for the data. The first column labels the classes by number and the second column indicates the class intervals. We then examine the 100 measurements of Table 3.5, keeping a tally of the number of measurements falling in each interval. The number of measurements falling in a given class interval is called the **class frequency.** These data are recorded in the third column of the frequency table. (See Table 3.6.)

class frequency

relative frequency

The **relative frequency** of a class is defined as the frequency of the class divided by the total number of measurements in the set (total frequency). Thus, if we let f_i denote the frequency for class i and let n denote the total number of

TABLE 3.6
Frequency table for number of attached ticks

Class	Class Interval	Frequency f_i	Relative Frequency f_i/n
1	16.25–18.75	2	.02
2	18.75–21.25	7	.07
3	21.25–23.75	7	.07
4	23.75–26.25	14	.14
5	26.25–28.75	17	.17
6	28.75–31.25	24	.24
7	31.25–33.75	11	.11
8	33.75–36.25	11	.11
9	36.25–38.75	3	.03
10	38.75–41.25	3	.03
11	41.25–43.75	1	.01
Totals		$n = 100$	1.00

FIGURE 3.6(a)
Frequency histogram for the tick data of Table 3.6

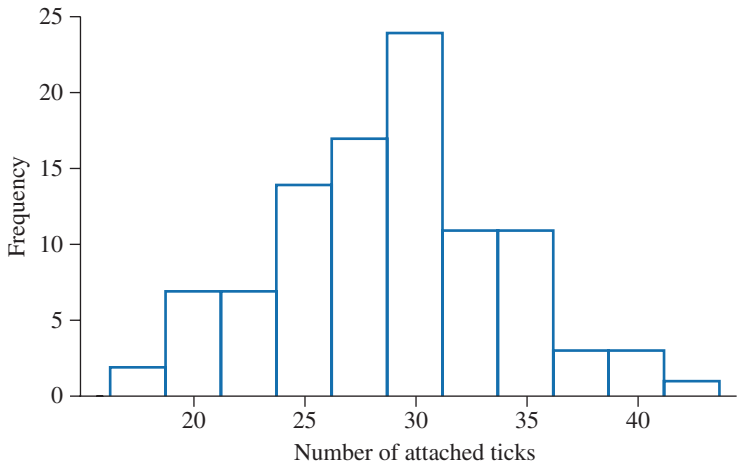
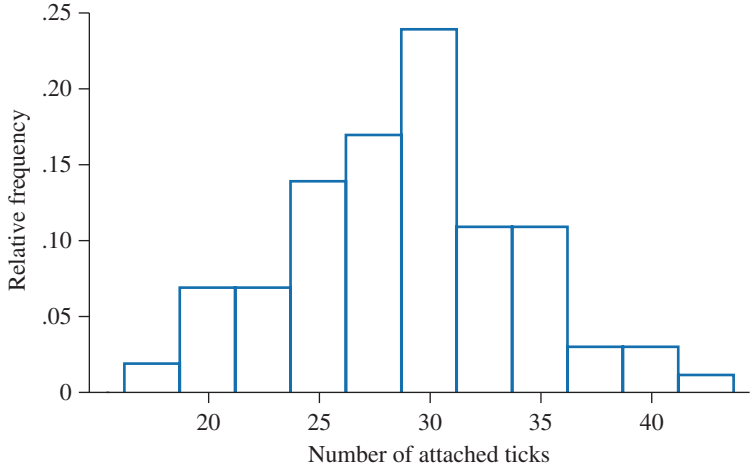


FIGURE 3.6(b)
Relative frequency histogram for the tick data of Table 3.6



measurements, the relative frequency for class i is f_i/n. The relative frequencies for all the classes are listed in the fourth column of Table 3.6.

The data of Table 3.5 have been organized into a frequency table, which can now be used to construct a *frequency histogram* or a *relative frequency histogram.* To construct a frequency histogram, draw two axes: a horizontal axis labeled with the class intervals and a vertical axis labeled with the frequencies. Then construct a rectangle over each class interval with a height equal to the number of measurements falling in a given subinterval. The frequency histogram for the data of Table 3.6 is shown in Figure 3.6(a).

The relative frequency histogram is constructed in much the same way as a frequency histogram. In the relative frequency histogram, however, the vertical axis is labeled as relative frequency, and a rectangle is constructed over each class interval with a height equal to the class relative frequency (the fourth column of Table 3.6). The relative frequency histogram for the data of Table 3.6 is shown in Figure 3.6(b). Clearly, the two histograms of Figures 3.6(a) and (b) are of the same shape and would be identical if the vertical axes were equivalent. We will frequently refer to either one as simply a **histogram.**

histogram

There are several comments that should be made concerning histograms. First, the distinction between bar charts and histograms is based on the distinction

between *qualitative* and *quantitative* variables. Values of qualitative variables vary in kind but not degree and hence are not measurements. For example, the variable political party affiliation can be categorized as Republican, Democrat, or other, and although we could label the categories as one, two, and three, these values are only codes and have no quantitative interpretation. In contrast, quantitative variables have actual units of measure. For example, the variable yield (in bushels) per acre of corn can assume specific values. *Pie charts and bar charts are used to display frequency data from qualitative variables; histograms are appropriate for displaying frequency data for quantitative variables.*

Second, the histogram is the most important graphical technique we will present because of the role it plays in statistical inference, a subject we will discuss in later chapters. Third, if we had an extremely large set of measurements, and if we constructed a histogram using many class intervals, each with a very narrow width, the histogram for the set of measurements would be, for all practical purposes, a smooth curve. Fourth, the fraction of the total number of measurements in an interval is equal to the fraction of the total area under the histogram over the interval.

For example, suppose we consider those intervals having cows with fewer numbers of ticks than the average under the previously used repellent—that is, the intervals containing cows having a number of attached ticks less than 34. From Table 3.6, we observe that exactly 82 of the 100 cows had fewer than 34 attached ticks. Thus, the proportion of the total measurements falling in those intervals—82/100 = .82—is equal to the proportion of the total area under the histogram over those intervals.

probability

Fifth, if a single measurement is selected at random from the set of sample measurements, the chance, or **probability,** that the selected measurement lies in a particular interval is equal to the fraction of the total number of sample measurements falling in that interval. This same fraction is used to estimate the probability that a measurement selected from the population lies in the interval of interest. For example, from the sample data of Table 3.5, the chance or probability of selecting a cow with less than 34 attached ticks is .82. The value .82 is an approximation of the proportion of all cows treated with the new repellant that would have fewer than 34 attached ticks after exposure to a population similar to that used in the study. In Chapters 5 and 6, we will introduce the process by which we can make a statement of our certainty that the new repellant is a significant improvement over the old repellant.

Because of the arbitrariness in the choice of number of intervals, starting value, and length of intervals, histograms can be made to take on different shapes for the same set of data, especially for small data sets. Histograms are most useful for describing data sets when the number of data points is fairly large—say, 50 or more. In Figures 3.7(a)–(d), a set of histograms for the tick data constructed using 5, 9, 13, and 18 class intervals illustrates the problems that can be encountered in attempting to construct a histogram. These graphs were obtained using the Minitab software program.

When the number of data points is relatively small and the number of intervals is large, the histogram fluctuates too much—that is, responds to a very few data values; see Figure 3.7(d). This results in a graph that is not a realistic depiction of the histogram for the whole population. When the number of class intervals is too small, most of the patterns or trends in the data are not displayed; see Figure 3.7(a). In the set of graphs in Figure 3.7, the histogram with 13 class intervals appears to be the most appropriate graph.

Finally, because we use proportions rather than frequencies in a relative frequency histogram, we can compare two different samples (or populations) by

FIGURE 3.7(a)
Relative frequency histogram for tick data (5 intervals)

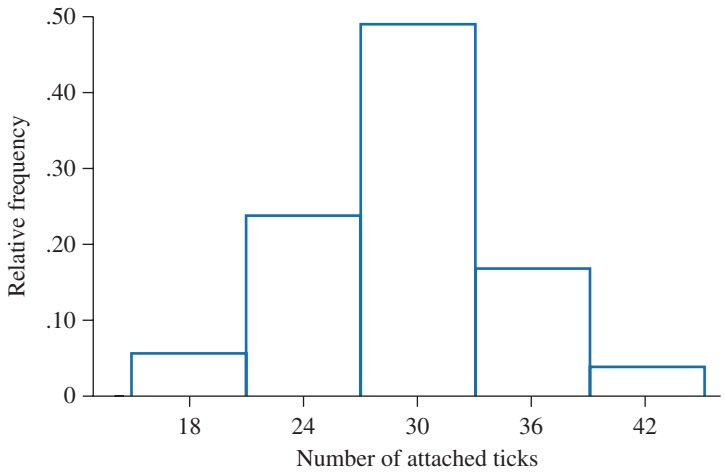


FIGURE 3.7(b)
Relative frequency histogram for tick data (9 intervals)

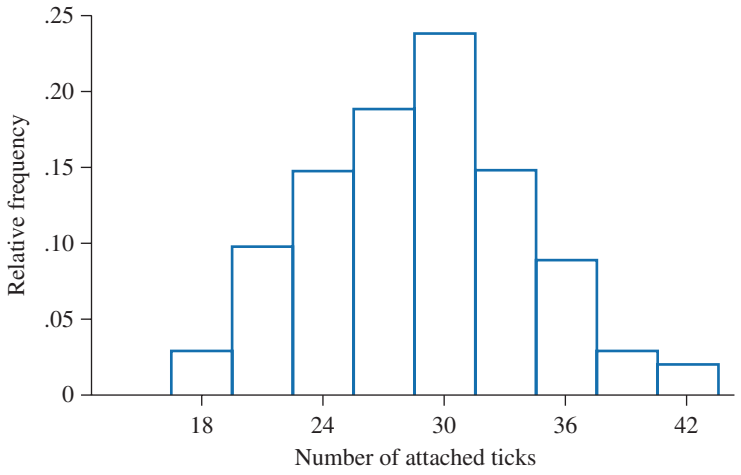


FIGURE 3.7(c)
Relative frequency histogram for tick data (13 intervals)

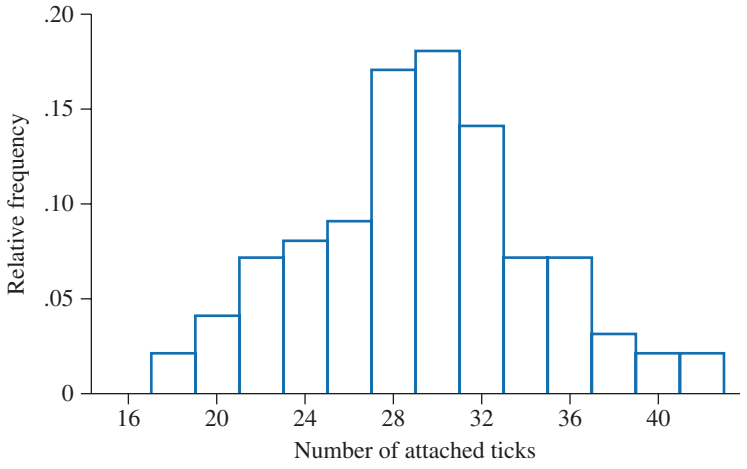
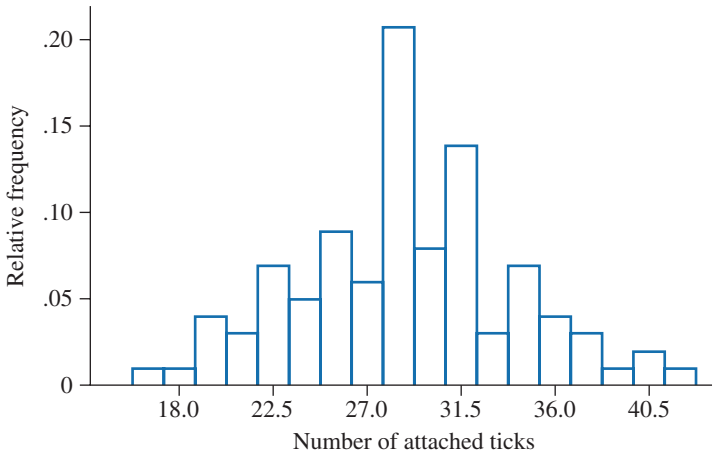


FIGURE 3.7(d)
Relative frequency histogram for tick data (18 intervals)



examining their relative frequency histograms even if the samples (populations) are of different sizes. When describing relative frequency histograms and comparing the plots from a number of samples, we examine the overall shape in the histogram. Figure 3.8 depicts many of the common shapes for relative frequency histograms.

unimodal

bimodal

A histogram with one major peak is called **unimodal;** see Figures 3.8(b), (c), and (d). When the histogram has two major peaks, such as in Figures 3.8(e) and (f), we state that the histogram is **bimodal.** In many instances, bimodal histograms are an indication that the sampled data are in fact from two distinct populations. Finally, when every interval has essentially the same number of observations, the histogram is called a **uniform** histogram; see Figure 3.8(a).

uniform

symmetric

A histogram is **symmetric** in shape if the right and left sides have essentially the same shape. Thus, Figures 3.8(a), (b), and (e) have symmetric shapes. When the right side of the histogram, containing the larger half of the observations in the data, extends a greater distance than the left side, the histogram is referred to as **skewed to the right;** see Figure 3.8(c). The histogram is **skewed to the left** when its left side extends a much larger distance than the right side; see Figure 3.8(d). We will see later in the text that knowing the shape of the distribution will help us choose the appropriate measures to summarize the data (Sections 3.4–3.7) and the methods for analyzing the data (Chapter 5 and beyond).

skewed to the right

skewed to the left

exploratory data analysis

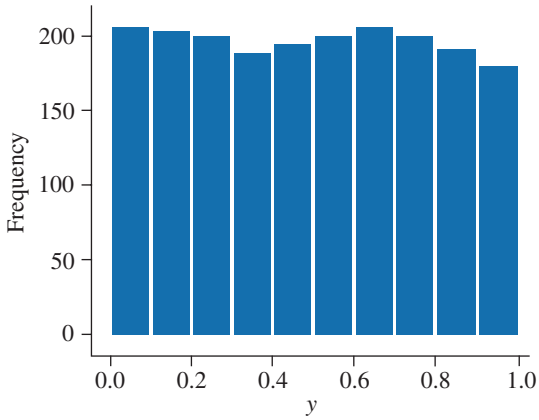
The next graphical technique presented in this section is a display technique taken from an area of statistics called **exploratory data analysis (EDA).** Professor John Tukey (1977) has been the leading proponent of this practical philosophy of data analysis aimed at exploring and understanding data.

stem-and-leaf plot

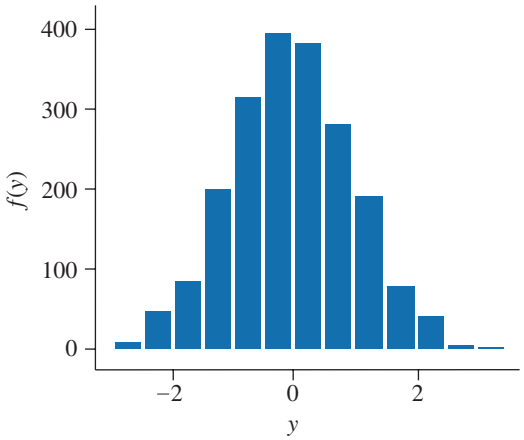
The **stem-and-leaf plot** is a clever, simple device for constructing a histogramlike picture of a frequency distribution. It allows us to use the information contained in a frequency distribution to show the range of scores, where the scores are concentrated, the shape of the distribution, whether there are any specific values or scores not represented, and whether there are any stray or extreme scores. The stem-and-leaf plot does not follow the organization principles stated previously for histograms. We will use the data shown in Table 3.7 to illustrate how to construct a stem-and-leaf plot.

The data in Table 3.7 are the maximum ozone readings (in parts per billion (ppb)) taken on 80 summer days in a large city. The readings are either two- or three-digit numbers. We will use the first digit of the two-digit numbers and the first two digits of the three-digit numbers as the stem number (see Figure 3.9) and the

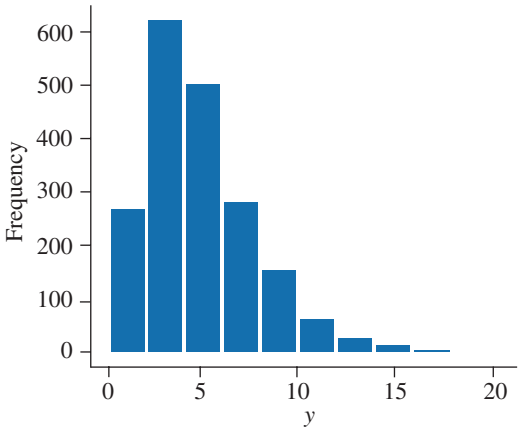
FIGURE 3.8 Some common shapes of distributions



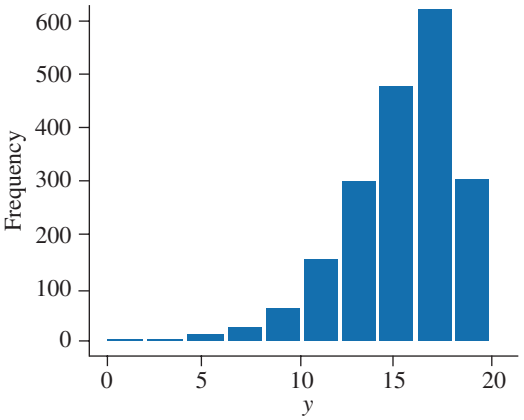
(a) Uniform distribution



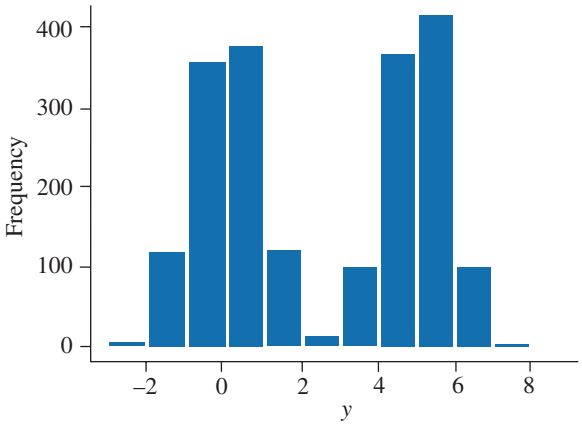
(b) Symmetric, unimodal (normal) distribution



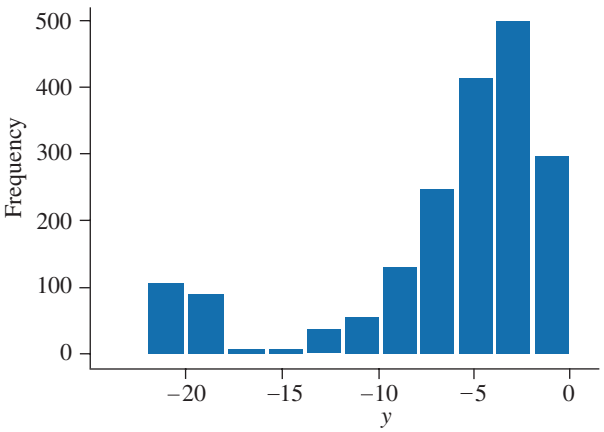
(c) Right-skewed distribution



(d) Left-skewed distribution



(e) Bimodal distribution



(f) Bimodal distribution skewed to left

TABLE 3.7
Maximum ozone readings (ppb)

60	61	61	64	64	64	64	66	66	68
68	68	69	71	71	71	71	71	71	72
72	73	75	75	80	80	80	80	80	80
82	82	83	85	86	86	87	87	87	89
91	92	94	94	98	99	99	100	101	103
103	103	108	111	113	113	114	118	119	119
122	122	124	124	124	125	125	131	133	134
136	141	142	143	146	150	152	155	169	169

FIGURE 3.9
Stem-and-leaf plot for maximum ozone readings (ppb) of Table 3.7

```
 6 0114444
 6 668889
 7 111111223
 7 55
 8 000000223
 8 5667779
 9 1244
 9 899
10 01333
10 8
11 1334
11 899
12 22444
12 55
13 134
13 6
14 123
14 6
15 02
15 5
16
16 99
```

remaining digits as the leaf number. For example, one of the readings was 85. Thus, 8 will be recorded as the stem number and 5 as the leaf number. A second maximum ozone reading was 111. Thus, 11 will be recorded as the stem number and 1 as the leaf number. If our data had been recorded in different units and resulted in, say, six-digit numbers such as 104,328, we might use the first two digits as stem numbers, use the second two digits as leaf numbers, and ignore the last two digits. This would result in some loss of information but would produce a much more useful graph.

For the data on maximum ozone readings, the smallest reading was 60 and the largest was 169. Thus, the stem numbers will be 6, 7, 8, . . . , 15, 16. In the same way that a class interval determines where a measurement is placed in a frequency table, the leading digits (stem of a measurement) determine the row in which a measurement is placed in a stem-and-leaf graph. The trailing digits for a measurement are then written in the appropriate row. In this way, each measurement is recorded in the stem-and-leaf plot, as in Figure 3.9 for the ozone data. The stem-and-leaf plot in Figure 3.9 was obtained using Minitab. Note that each of the stems is repeated twice, with leaf digits split into two groups: 0 to 4 and 5 to 9.

We can see that each stem defines a class interval and that the limits of each interval are the largest and smallest possible scores for the class. The values represented by each leaf must be between the lower and upper limits of the interval.

Note that a stem-and-leaf plot is a graph that looks much like a histogram turned sideways, as in Figure 3.9. The plot can be made a bit more useful by ordering the data (leaves) within a row (stem) from lowest to highest as we did in

Figure 3.9. The advantage of such a graph over the histogram is that it reflects not only the frequencies, concentration(s) of scores, and shapes of the distribution but also the actual scores. The disadvantage is that for large data sets, the stem-and-leaf plot can be more unwieldy than the histogram.

Guidelines for Constructing Stem-and-Leaf Plots

1. Split each score or value into two sets of digits. The first or leading set of digits is the stem and the second or trailing set of digits is the leaf.
2. List all possible stem digits from lowest to highest.
3. For each score in the mass of data, write the leaf values on the line labeled by the appropriate stem number.
4. If the display looks too cramped and narrow, stretch the display by using two lines per stem so that, for example, leaf digits 0, 1, 2, 3, and 4 are placed on the first line of the stem and leaf digits 5, 6, 7, 8, and 9 are placed on the second line.
5. If too many digits are present, such as in a six- or seven-digit score, drop the right-most trailing digit(s) to maximize the clarity of the display.
6. The rules for developing a stem-and-leaf plot are somewhat different from the rules governing the establishment of class intervals for the traditional frequency distribution and for a variety of other procedures that we will consider in later sections of the text. Class intervals for stem-and-leaf plots are, then, in a sense slightly atypical.

The following data display and stem-and-leaf plot (Figure 3.10) are obtained from Minitab. The data consist of the number of employees in the wholesale and retail trade industries in Wisconsin measured each month for a 5-year period.

```
Data Display

Trade
    322   317   319   323   327   328   325   326   330   334
    337   341   322   318   320   326   332   334   335   336
    335   338   342   348   330   325   329   337   345   350
    351   354   355   357   362   368   348   345   349   355
    362   367   366   370   371   375   380   385   361   354
    357   367   376   381   381   383   384   387   392   396
```

Note that most of the stems are repeated twice, with the leaf digits split into two groups: 0 to 4 and 5 to 9.

The last graphical technique to be presented in this section deals with how certain variables change over time. For macroeconomic data such as disposable income and microeconomic data such as weekly sales data of one particular product at one particular store, plots of data over time are fundamental to business management. Similarly, social researchers are often interested in showing how variables change over time. They might be interested in changes with time in attitudes toward various racial and ethnic groups, changes in the rate of savings in the United States, or changes in crime rates for various cities. A pictorial method of

FIGURE 3.10
Character stem-and-leaf display for trade data

```
Stem-and-leaf of Trade     N = 60
Leaf Unit = 1.0

       31 789
       32 0223
       32 5666789
       33 00244
       33 556778
       34 12
       34 55889
       35 0144
       35 5577
       36 122
       36 6778
       37 01
       37 56
       38 01134
       38 57
       39 2
       39 6
```

time series

presenting changes in a variable over time is called a **time series.** Figure 3.11 is a time series showing the number of homicides, forcible rapes, robberies, and aggravated assaults included in the ***Uniform Crime Reports*** of the FBI.

Usually, the time points are labeled chronologically across the horizontal axis (abscissa), and the numerical values (frequencies, percentages, rates, etc.) of the variable of interest are labeled along the vertical axis (ordinate). Time can be measured in days, months, years, or whichever unit is most appropriate. As a rule of thumb, a time series should consist of no fewer than four or five time points; typically, these time points are equally spaced. Many more time points than this are desirable, though, in order to show a more complete picture of changes in a variable over time.

How we display the time axis in a time series frequently depends on the time intervals at which data are available. For example, the U.S. Census Bureau reports average family income in the United States only on a yearly basis. When information about a variable of interest is available in different units of time, we must decide which unit or units are most appropriate for the research. In an election

FIGURE 3.11
Total violent crimes in the United States, 1983–2012
Source: ***Uniform Crime Reports.***

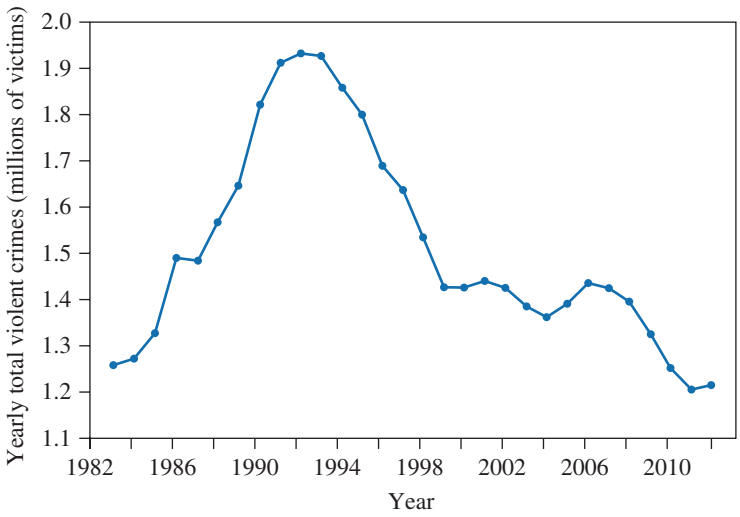
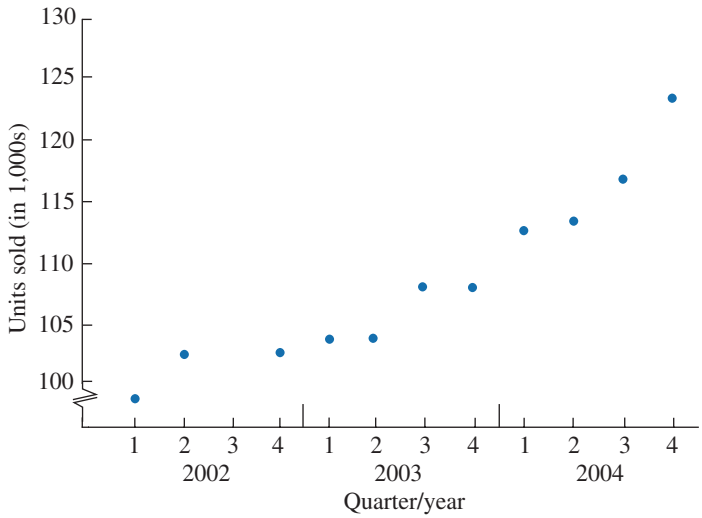


FIGURE 3.12
Quarterly sales (in thousands)

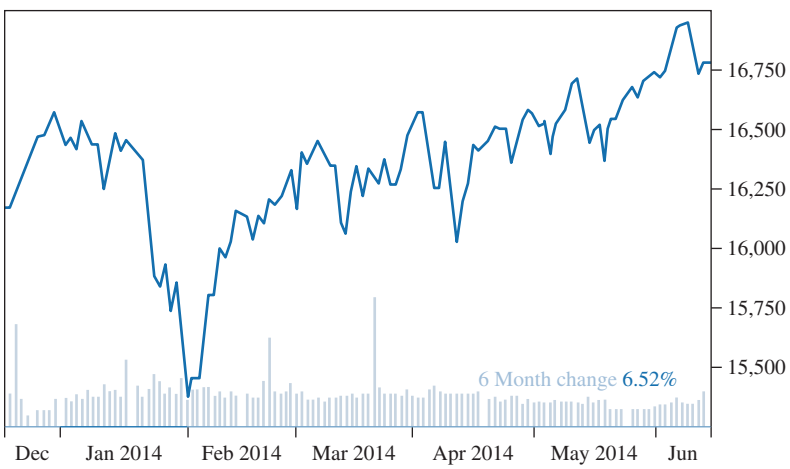


year, a political scientist would most likely examine weekly or monthly changes in candidate preferences among registered voters. On the other hand, a manufacturer of machine-tool equipment might keep track of sales (in dollars and number of units) on a monthly, quarterly, and yearly basis. Figure 3.12 shows the quarterly sales (in thousands of units) of a machine-tool product over 3 years. Note that from this time series it is clear that the company has experienced a gradual but steady growth in the number of units over the 3 years.

Time-series plots are useful for examining general trends and seasonal or cyclic patterns. For example, the "Money and Investing" section of the ***Wall Street Journal*** gives the daily workday values for the Dow Jones Industrials Averages. Figure 3.13 displays the daily Dow Jones Industrial Average for the period from mid-December 2013 through mid-June 2014. Exercise 3.58 provides the details on how the Dow Jones Industrial Average is computed. The plot reveals a sharp decline in values from mid-January to the beginning of February. This decline is followed by a steady increase through mid-June 2014. However, there are just enough daily decreases in the Dow values to keep investors nervous. In order to detect seasonal or cyclical patterns in a time series, there must be daily values recorded over a large number of years.

FIGURE 3.13
Time-series plot of the Dow Jones Average, mid-December 2013 to Mid-June 2014

Source: ***Wall Street Journal.***



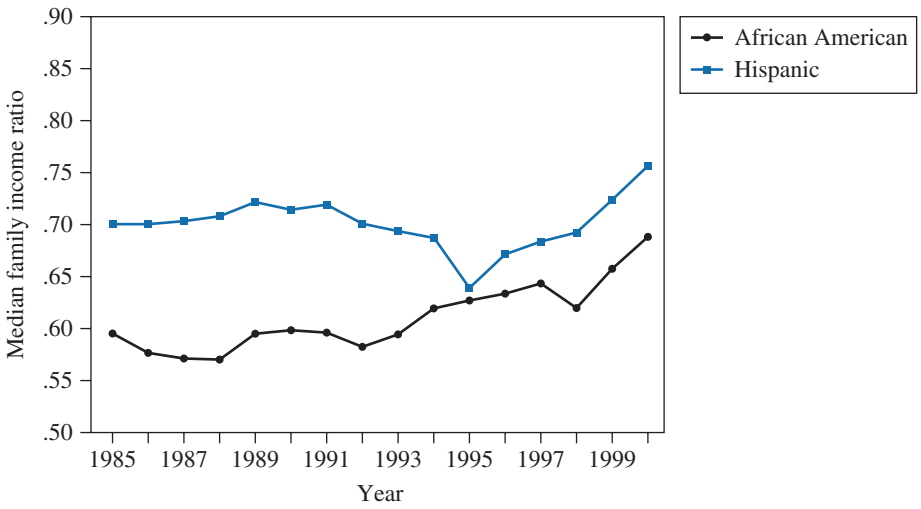


FIGURE 3.14
Ratio of African American and Hispanic median family incomes to Anglo-American median family income.
Source: U.S. Census Bureau.

Sometimes it is important to compare trends over time in a variable for two or more groups. Figure 3.14 reports the values of two ratios from 1985 to 2000: the ratio of the median family income of African Americans to the median family income of Anglo-Americans and the ratio of the median family income of Hispanics to the median family income of Anglo-Americans.

Median family income represents the income amount that divides family incomes into two groups—the top half and the bottom half. For example, in 1987, the median family income for African Americans was $18,098, meaning that 50% of all African American families had incomes above $18,098 and 50% had incomes below $18,098. The median, one of several measures of central tendency, is discussed more fully later in this chapter.

Figure 3.14 shows that the ratio of African American to Anglo-American family income and the ratio of Hispanic to Anglo-American family income remained fairly constant from 1985 to 1991. From 1995 to 2000, there was an increase in both ratios and a narrowing of the difference between the ratio of African American family income and the ratio of Hispanic family income. We can interpret this trend to mean that the income of African American and Hispanic families has generally increased relative to the income of Anglo-American families.

Sometimes information is not available in equal time intervals. For example, polling organizations such as Gallup or the National Opinion Research Center do not necessarily ask the American public the same questions about their attitudes or behavior on a yearly basis. Sometimes there is a time gap of more than 2 years before a question is asked again.

When information is not available in equal time intervals, it is important for the interval width between time points (the horizontal axis) to reflect this fact. If, for example, a social researcher is plotting values of a variable for 1995, 1996, 1997, and 2000, the interval width between 1997 and 2000 on the horizontal axis should be three times the width of that between the other years. If these interval widths were spaced evenly, the resulting trend line could be seriously misleading.

Before leaving graphical methods for describing data, there are several general guidelines that can be helpful in developing graphs with an impact. These guidelines pay attention to the design and presentation techniques and should help you make better, more informative graphs.

General Guidelines for Developing Successful Graphics

1. Before constructing a graph, set your priorities. What messages should the viewer get?
2. Choose the type of graph (pie chart, bar graph, histogram, and so on).
3. Pay attention to the title. One of the most important aspects of a graph is its title. The title should immediately inform the viewer of the point of the graph and draw the eye toward the most important elements of the graph.
4. Fight the urge to use many type sizes, styles, and colors. The indiscriminate and excessive use of different type sizes, styles, and colors will confuse the viewer. Generally, we recommend using only two typefaces; color changes and italics should be used in only one or two places.
5. Convey the tone of your graph by using colors and patterns. Intense, warm colors (yellows, oranges, reds) are more dramatic than the blues and purples and help to stimulate enthusiasm by the viewer. On the other hand, pastels (particularly grays) convey a conservative, businesslike tone. Similarly, simple patterns convey a conservative tone, whereas busier patterns stimulate more excitement.
6. Don't underestimate the effectiveness of a simple, straightforward graph.

3.4 Describing Data on a Single Variable: Measures of Central Tendency

Numerical descriptive measures are commonly used to convey a mental image of pictures, objects, and other phenomena. There are two main reasons for this. First, graphical descriptive measures are inappropriate for statistical inference because it is difficult to describe the similarity of a sample frequency histogram and the corresponding population frequency histogram. The second reason for using numerical descriptive measures is one of expediency—we never seem to carry the appropriate graphs or histograms with us and so must resort to our powers of verbal communication to convey the appropriate picture. We seek several numbers, called *numerical descriptive measures,* that will create a mental picture of the frequency distribution for a set of measurements.

central tendency
variability

The two most common numerical descriptive measures are measures of **central tendency** and measures of **variability;** that is, we seek to describe the center of the distribution of measurements and also how the measurements vary about the center of the distribution. We will draw a distinction between numerical descriptive measures for a population, called **parameters,** and numerical descriptive measures for a sample, called **statistics.** In problems requiring statistical inference, we will not be able to calculate values for various parameters, but we will be able to compute corresponding statistics from the sample and use these quantities to estimate the corresponding population parameters.

parameters
statistics

In this section, we will consider various measures of central tendency, followed in Section 3.5 by a discussion of measures of variability.

mode

The first measure of central tendency we consider is the **mode.**

DEFINITION 3.1

The **mode** of a set of measurements is defined to be the measurement that occurs most often (with the highest frequency).

We illustrate the use and determination of the mode in an example.

EXAMPLE 3.1

A consumer investigator is interested in the differences in the selling prices of a new popular compact automobile at various dealers in a 100-mile radius of Houston, Texas. She asks for a quote from 25 dealers for this car with exactly the same options. The selling prices (in 1,000s) are given here.

26.6	25.3	23.8	24.0	27.5
21.1	25.9	22.6	23.8	25.1
22.6	27.5	26.8	23.4	27.5
20.8	20.4	22.4	27.5	23.7
22.2	23.8	23.2	28.7	27.5

Determine the modal selling price.

Solution For these data, the price 23.8 occurred three times in the sample, but the price 27.5 occurred five times. Because no other value occurred more than once, we would state the data had a modal selling price of $27,500. ■

Identification of the mode for Example 3.1 was quite easy because we were able to count the number of times each measurement occurred. When dealing with grouped data—data presented in the form of a frequency table—we can define the modal interval to be the class interval with the highest frequency. However, because we would not know the actual measurements but only how many measurements fall into each interval, the mode is taken as the midpoint of the modal interval; it is an approximation to the mode of the actual sample measurements.

The mode is also commonly used as a measure of popularity that reflects central tendency or opinion. For example, we might talk about the most preferred stock, the most preferred model of washing machine, or the most popular candidate. In each case, we would be referring to the mode of the distribution. In Figure 3.8 of the previous section, frequency histograms (b), (c), and (d) had a single mode, with that mode located at the center of the class having the highest frequency. Thus, the modes would be $-.25$ for histogram (b), 3 for histogram (c), and 17 for histogram (d). It should be noted that some distributions have more than one measurement that occurs with the highest frequency. Thus, we might encounter distributions that are bimodal, trimodal, and so on. In Figure 3.8, histogram (e) is essentially bimodal, with nearly equal peaks at $y = 0.5$ and $y = 5.5$.

median

The second measure of central tendency we consider is the **median.**

DEFINITION 3.2

The **median** of a set of measurements is defined to be the middle value when the measurements are arranged from lowest to highest.

The median is most often used to measure the midpoint of a large set of measurements. For example, we may read about the median wage increase won by union members, the median age of persons receiving Social Security benefits, and the median weight of cattle prior to slaughter during a given month. Each of these situations involves a large set of measurements, and the median would reflect the

central value of the data—that is, the value that divides the set of measurements into two groups, with an equal number of measurements in each group.

However, we may use the definition of median for small sets of measurements by using the following convention: The median for an even number of measurements is the average of the two middle values when the measurements are arranged from lowest to highest. When there are an odd number of measurements, the median is still the middle value. Thus, whether there are an even or odd number of measurements, there are an equal number of measurements above and below the median.

EXAMPLE 3.2

After the third-grade classes in a school district received low overall scores on a statewide reading test, a supplemental reading program was implemented in order to provide extra help to those students who were below expectations with respect to their reading proficiency. Six months after implementing the program, the 10 third-grade classes in the district were reexamined. For each of the 10 schools, the percentage of students reading above the statewide standard was determined. These data are shown here.

95 86 78 90 62 73 89 92 84 76

Determine the median percentage of the 10 schools.

Solution First, we must arrange the percentages in order of magnitude.

62 73 76 78 84 86 89 90 92 95

Because there are an even number of measurements, the median is the average of the two midpoint scores.

$$\text{median} = \frac{84 + 86}{2} = 85 \blacksquare$$

EXAMPLE 3.3

An experiment was conducted to measure the effectiveness of a new procedure for pruning grapes. Each of 13 workers was assigned the task of pruning an acre of grapes. The productivity, measured in worker-hours/acre, was recorded for each person.

4.4 4.9 4.2 4.4 4.8 4.9 4.8 4.5 4.3 4.8 4.7 4.4 4.2

Determine the mode and median productivity for the group.

Solution First, arrange the measurements in order of magnitude:

4.2 4.2 4.3 4.4 4.4 4.4 4.5 4.7 4.8 4.8 4.8 4.9 4.9

For these data, we have two measurements appearing three times each. Hence, the data are bimodal, with modes of 4.4 and 4.8. The median for the odd number of measurements is the middle score, 4.5. ■

grouped data median

The **median for grouped data** is slightly more difficult to compute. Because the actual values of the measurements are unknown, we know that the median occurs in a particular class interval, but we do not know where to locate the median within

the interval. If we assume that the measurements are spread evenly throughout the interval, we get the following result. Let

L = lower class limit of the interval that contains the median

n = total frequency

cf_b = the sum of frequencies (cumulative frequency) for all classes before the median class

f_m = frequency of the class interval containing the median

w = interval width

Then, for grouped data,

$$\text{median} = L + \frac{w}{f_m}(.5n - cf_b)$$

The next example illustrates how to find the median for grouped data.

EXAMPLE 3.4

Table 3.8 is a repeat of the frequency table (Table 3.6) with some additional columns for the tick data of Table 3.5. Compute the median number of ticks per cow for these data.

TABLE 3.8
Frequency table for number of attached ticks, Table 3.5

Class	Class Interval	f_i	Cumulative f_i	f_i/n	Cumulative f_i/n
1	16.25–18.75	2	2	.02	.02
2	18.75–21.25	7	9	.07	.09
3	21.25–23.75	7	16	.07	.16
4	23.75–26.25	14	30	.14	.30
5	26.25–28.75	17	47	.17	.47
6	28.75–31.25	24	71	.24	.71
7	31.25–33.75	11	82	.11	.82
8	33.75–36.25	11	93	.11	.93
9	36.25–38.75	3	96	.03	.96
10	38.75–41.25	3	99	.03	.99
11	41.25–43.75	1	100	.01	1.00

Solution Let the cumulative relative frequency for class j equal the sum of the relative frequencies for class 1 through class j. To determine the interval that contains the median, we must find the first interval for which the cumulative relative frequency exceeds .50. This interval is the one containing the median. For these data, the interval from 28.75 to 31.25 is the first interval for which the cumulative relative frequency exceeds .50, as shown in Table 3.8, Class 6. So this interval contains the median. Then

$L = 28.75 \quad f_m = 24$

$n = 100 \quad w = 2.5$

$cf_b = 47$

and

$$\text{median} = L + \frac{w}{f_m}(.5n - cf_b) = 28.75 + \frac{2.5}{24}(50 - 47) = 29.06 \blacksquare$$

Note that the value of the median from the ungrouped data of Table 3.5 is 29. Thus, the approximated value and the value from the ungrouped data are nearly equal. The difference between the two values for the sample median decreases as the number of class intervals increases.

mean

The third, and last, measure of central tendency we will discuss in this text is the arithmetic mean, known simply as the **mean.**

DEFINITION 3.3

The **arithmetic mean,** or **mean,** of a set of measurements is defined to be the sum of the measurements divided by the total number of measurements.

μ
$\bar{y}$

When people talk about an "average," they quite often are referring to the mean. It is the balancing point of the data set. Because of the important role that the mean will play in statistical inference in later chapters, we give special symbols to the population mean and the sample mean. The *population mean* is denoted by the Greek letter $\boldsymbol{\mu}$ (read "mu"), and the *sample mean* is denoted by the symbol $\bar{y}$ (read "y-bar"). As indicated in Chapter 1, a population of measurements is the complete set of measurements of interest to us; a sample of measurements is a subset of measurements selected from the population of interest. If we let $y_1, y_2, \ldots, y_n$ denote the measurements observed in a sample of size n, then the sample mean $\bar{y}$ can be written as

$$\bar{y} = \frac{\sum_i y_i}{n}$$

where the symbol appearing in the numerator, $\sum_i y_i$, is the notation used to designate a sum of n measurements, y_i:

$$\sum_i y_i = y_1 + y_2 + \cdots + y_n$$

The corresponding population mean is μ.

In most situations, we will not know the population mean; the sample will be used to make inferences about the corresponding unknown population mean. For example, the accounting department of a large department store chain is conducting an examination of its overdue accounts. The store has thousands of such accounts, which would yield a population of overdue values having a mean value, μ. The value of μ could be determined only by conducting a large-scale audit that would take several days to complete. The accounting department monitors the overdue accounts on a daily basis by taking a random sample of n overdue accounts and computing the sample mean, $\bar{y}$. The sample mean, $\bar{y}$, is then used as an estimate of the mean value, μ, of *all* overdue accounts for that day. The accuracy of the estimate and approaches for determining the appropriate sample size will be discussed in Chapter 5.

EXAMPLE 3.5

A sample of $n = 15$ overdue accounts in a large department store yields the following amounts due:

$55.20	$ 4.88	$271.95
18.06	180.29	365.29
28.16	399.11	807.80
44.14	97.47	9.98
61.61	56.89	82.73

a. Determine the mean amount due for the 15 accounts sampled.
b. If there are a total of 150 overdue accounts, use the sample mean to predict the total amount overdue for all 150 accounts.

Solution

a. The sample mean is computed as follows:

$$\bar{y} = \frac{\sum_i y_i}{15} = \frac{55.20 + 18.06 + \cdots + 82.73}{15} = \frac{2{,}483.56}{15} = \$165.57$$

b. From part (a), we found that the 15 accounts sampled averaged \$165.57 overdue. Using this information, we would predict, or estimate, the total amount overdue for the 150 accounts to be 150(165.57) = \$24,835.50. ■

The sample mean formula for grouped data is only slightly more complicated than the formula just presented for ungrouped data. In certain situations, the original data will be presented in a frequency table or a histogram. Thus, we will not know the individual sample measurements, only the interval to which a measurement is assigned. In this type of situation, the formula for the mean from the grouped data will be an approximation to the actual sample mean. Hence, when the sample measurements are known, the formula for ungrouped data should be used. If there are k class intervals and

y_i = midpoint of the ith class interval
f_i = frequency associated with the ith class interval
n = total number of measurements

then

$$\bar{y} \cong \frac{\sum_i f_i y_i}{n}$$

where ≅ denotes "is approximately equal to."

EXAMPLE 3.6

The data of Example 3.4 are reproduced in Table 3.9, along with three additional columns: y_i, $f_i y_i$, $f_i(y_i - \bar{y})^2$. These values will be needed in order to compute approximations to the sample mean and the sample standard deviation. Using the information in Table 3.9, compute an approximation to the sample mean for this set of grouped data.

TABLE 3.9
Class information for number of attached ticks

Class	Class Interval	f_i	y_i	$f_i y_i$	$f_i(y_i - \bar{y})^2$
1	16.25–18.75	2	17.5	35.0	258.781
2	18.75–21.25	7	20.0	140.0	551.359
3	21.25–23.75	7	22.5	157.5	284.484
4	23.75–26.25	14	25.0	350.0	210.219
5	26.25–28.75	17	27.5	467.5	32.141
6	28.75–31.25	24	30.0	720.0	30.375
7	31.25–33.75	11	32.5	357.5	144.547
8	33.75–36.25	11	35.0	385.0	412.672
9	36.25–38.75	3	37.5	112.5	223.172
10	38.75–41.25	3	40.0	120.0	371.297
11	41.25–43.75	1	42.5	42.5	185.641
	Totals	100		2,887.5	2,704.688

Solution After adding the entries in the $f_i y_i$ column and substituting into the formula, we determine that an approximation to the sample mean is

$$\bar{y} \cong \frac{\sum_{i=1}^{11} f_i y_i}{100} = \frac{2{,}887.5}{100} = 28.875$$

Using the 100 values, from Table 3.5, the actual value of the sample mean is

$$\bar{y} = \frac{\sum_{i=1}^{100} y_i}{100} = \frac{2{,}881}{100} = 28.81 \blacksquare$$

Example 3.6 demonstrates that the approximation from the grouped data formula can be very close to the actual value. When the number of class intervals is relatively large, the approximation from the grouped data formula will be very close to the actual sample mean.

The mean is a useful measure of the central value of a set of measurements, but it is subject to distortion due to the presence of one or more extreme values in the set. In these situations, the extreme values (called **outliers**) pull the mean in the direction of the outliers to find the balancing point, thus distorting the mean as a measure of the central value. A variation of the mean, called a **trimmed mean,** drops the highest and lowest extreme values and averages the rest. For example, a 5% trimmed mean drops the highest 5% and the lowest 5% of the measurements and averages the rest. Similarly, a 10% trimmed mean drops the highest and the lowest 10% of the measurements and averages the rest. In Example 3.5, a 10% trimmed mean would drop the smallest and largest account, resulting in a mean of

outliers

trimmed mean

$$\bar{y} = \frac{2{,}483.56 - 4.88 - 807.8}{13} = \$128.53$$

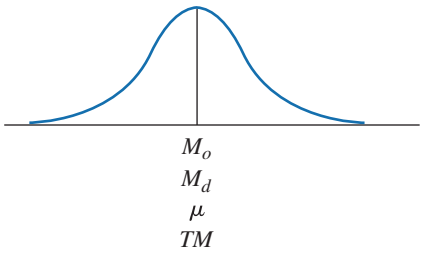
By trimming the data, we are able to reduce the impact of very large (or small) values on the mean and thus get a more reliable measure of the central value of the set. This will be particularly important when the sample mean is used to predict the corresponding population central value.

Note that in a limiting sense the median is a 50% trimmed mean. Thus, the median is often used in place of the mean when there are extreme values in the data set. In Example 3.5, the value $807.80 is considerably larger than the other values in the data set. This results in 10 of the 15 accounts having values less than the mean and only 5 having values larger than the mean. The median value for the 15 accounts is $61.61. There are 7 accounts less than the median and 7 accounts greater than the median. Thus, in selecting a typical overdue account, the median is a more appropriate value than the mean. However, if we want to estimate the total amount overdue in all 150 accounts, we would want to use the mean and not the median. When estimating the sum of all measurements in a population, we would not want to exclude the extremes in the sample. Suppose a sample contains a few extremely large values. If the extremes are trimmed, then the population sum will be grossly underestimated using the sample trimmed mean or sample median in place of the sample mean.

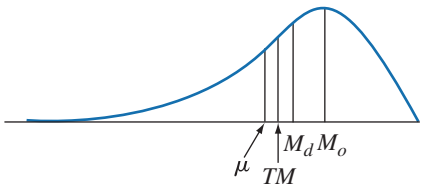
In this section, we discussed the mode, median, mean, and trimmed mean. How are these measures of central tendency related for a given set of measurements? The answer depends on the **skewness** of the data. If the distribution is mound-shaped and symmetrical about a single peak, the mode (M_o), median (M_d), mean (μ), and trimmed mean (TM) will all be the same. This is shown using a

skewness

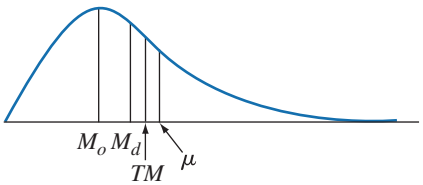
FIGURE 3.15
Relation among the mean μ, the trimmed mean TM, the median M_d, and the mode M_o



(a) A mound-shaped distribution



(b) A distribution skewed to the left



(c) A distribution skewed to the right

smooth curve and population quantities in Figure 3.15(a). If the distribution is skewed, having a long tail in one direction and a single peak, the mean is pulled in the direction of the tail; the median falls between the mode and the mean; and depending on the degree of trimming, the trimmed mean usually falls between the median and the mean. Figures 3.15(b) and (c) illustrate this for distributions skewed to the left and to the right.

The important thing to remember is that we are not restricted to using only one measure of central tendency. For some data sets, it will be necessary to use more than one of these measures to provide an accurate descriptive summary of central tendency for the data.

Major Characteristics of Each Measure of Central Tendency

Mode

1. It is the most frequent or probable measurement in the data set.
2. There can be more than one mode for a data set.
3. It is not influenced by extreme measurements.
4. Modes of subsets cannot be combined to determine the mode of the complete data set.
5. For grouped data, its value can change depending on the categories used.
6. It is applicable for both qualitative and quantitative data.

Median

1. It is the central value; 50% of the measurements lie above it and 50% fall below it.
2. There is only one median for a data set.

3. It is not influenced by extreme measurements.
4. Medians of subsets cannot be combined to determine the median of the complete data set.
5. For grouped data, its value is rather stable even when the data are organized into different categories.
6. It is applicable to quantitative data only.

Mean

1. It is the arithmetic average of the measurements in a data set.
2. There is only one mean for a data set.
3. Its value is influenced by extreme measurements; trimming can help to reduce the degree of influence.
4. Means of subsets can be combined to determine the mean of the complete data set.
5. It is applicable to quantitative data only.

Measures of central tendency do not provide a complete mental picture of the frequency distribution for a set of measurements. In addition to determining the center of the distribution, we must have some measure of the spread of the data. In the next section, we discuss measures of variability, or dispersion.

3.5 Describing Data on a Single Variable: Measures of Variability

It is not sufficient to describe a data set using only measures of central tendency, such as the mean or the median. For example, suppose we are monitoring the production of plastic sheets that have a nominal thickness of 3 mm. If we randomly select 100 sheets from the daily output of the plant and find that the average thickness of the 100 sheets is 3 mm, does this indicate that all 100 sheets have the desired thickness of 3 mm? We may have a situation in which 50 sheets have a thickness of 1 mm and the remaining 50 sheets have a thickness of 5 mm. This would result in an average thickness of 3 mm, but none of the 100 sheets would have a thickness close to the specified 3 mm. Thus, we need to determine how dispersed the sheet thicknesses are about the mean of 3 mm.

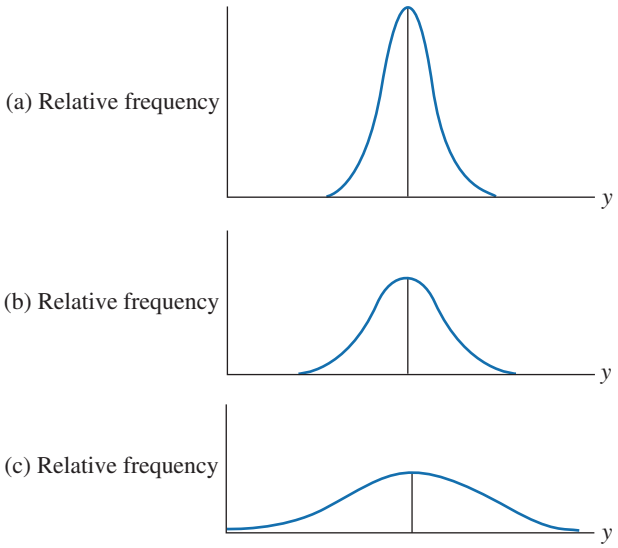
variability

Graphically, we can observe the need for some measure of variability by examining the relative frequency histograms of Figure 3.16. All the histograms have the same mean, but each has a different spread, or **variability,** about the mean. For illustration, we have shown the histograms as smooth curves. Suppose the three histograms represent the amount of PCB (ppb) found in a large number of 1-liter samples taken from three lakes that are close to chemical plants. The average amount of PCB, μ, in a 1-liter sample is the same for all three lakes. However, the variabilities in the PCB quantities are considerably different. Thus, the lake with the PCB quantities depicted in histogram (a) would have fewer samples containing very small or large quantities of PCB as compared to the lake with PCB values depicted in histogram (c). Knowing only the mean PCB quantity in the three lakes would mislead the investigator concerning the level of PCB present in all three lakes.

range

The simplest but least useful measure of data variation is the **range,** which we alluded to in Section 3.2. We now present its definition.

FIGURE 3.16
Relative frequency histograms with different variabilities but the same mean



DEFINITION 3.4

The **range** of a set of measurements is defined to be the difference between the largest and the smallest measurements of the set.

EXAMPLE 3.7

Determine the range of the 15 overdue accounts of Example 3.5.

Solution The smallest measurement is \$4.88 and the largest is \$807.80. Hence, the range is

$$\$807.80 - \$4.88 = \$802.92 \blacksquare$$

grouped data

For **grouped data,** because we do not know the individual measurements, the **range** is taken to be the difference between the upper limit of the last interval and the lower limit of the first interval.

Although the range is easy to compute, it is sensitive to outliers because it depends on the most extreme values. It does not give much information about the pattern of variability. Referring to the situation described in Example 3.5, if in the current budget period the 15 overdue accounts consisted of 10 accounts having a value of \$4.88, 3 accounts of \$807.80, and 2 accounts of \$5.68, then the mean value would be \$165.57 and the range would be \$802.92. The mean and range would be identical to the mean and range calculated for the data of Example 3.5. However, the data in the current budget period are more spread out about the mean than the data in the earlier budget period. What we seek is a measure of variability that discriminates between data sets having different degrees of concentration of the data about the mean.

percentiles

A second measure of variability involves the use of **percentiles.**

DEFINITION 3.5

The ***p*th percentile** of a set of n measurements arranged in order of magnitude is that value that has at most p% of the measurements below it and at most $(100 - p)$% above it.

FIGURE 3.17
The 60th percentile of a set of measurements

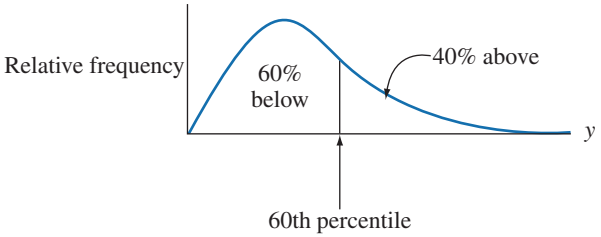
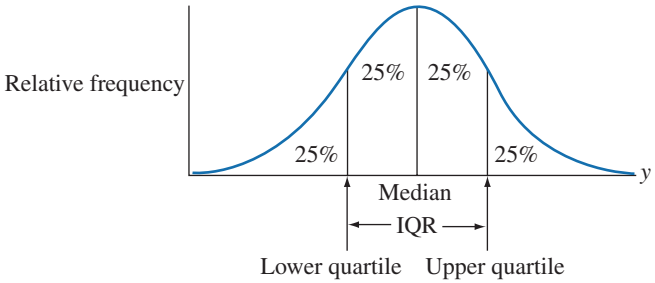


FIGURE 3.18
Quartiles of a distribution



For example, Figure 3.17 illustrates the 60th percentile of a set of measurements. Percentiles are frequently used to describe the results of achievement test scores and the ranking of a person in comparison to the rest of the people taking an examination. Specific percentiles of interest are the 25th, 50th, and 75th percentiles, often called the *lower quartile,* the *middle quartile* (median), and the *upper quartile,* respectively (see Figure 3.18).

The computation of percentiles is accomplished as follows: Each data value corresponds to a percentile for the percentage of the data values that are less than or equal to it. Let $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ denote the ordered observations for a data set; that is,

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$$

The ith ordered observation, $y_{(i)}$, corresponds to the $100(i - .5)/n$ percentile. We use this formula in place of assigning the percentile $100i/n$ so that we avoid assigning the 100th percentile to $y_{(n)}$, which would imply that the largest possible data value in the population was observed in the data set, an unlikely happening. For example, a study of serum total cholesterol (mg/l) levels recorded the levels given in Table 3.10 for 20 adult patients. Thus, each ordered observation is a data percentile corresponding to a multiple of the fraction $100(i - .5)/n = 100(2i - 1)/2n = 100(2i - 1)/40$.

The 22.5th percentile is 152 (mg/l). Thus, 22.5% of persons in the study have a serum cholesterol less than or equal to 152. Also, the median of the above data set, which is the 50th percentile, is halfway between 192 and 201; that is, median $= (192 + 201)/2 = 196.5$. Thus, approximately half of the persons in the study have a serum cholesterol level less than 196.5 and half have a level greater than 196.5.

When dealing with large data sets, the percentiles are generalized to quantiles, where a quantile, denoted $Q(u)$, is a number that divides a sample of n data values into two groups so that the specified fraction u of the data values is less than or equal to the value of the quantile, $Q(u)$. Plots of the quantiles $Q(u)$ versus the data fraction u provide a method of obtaining estimated quantiles for the

TABLE 3.10
Serum cholesterol levels

Observation (j)	Cholesterol (mg/l)	Percentile
1	133	2.5
2	137	7.5
3	148	12.5
4	149	17.5
5	152	22.5
6	167	27.5
7	174	32.5
8	179	37.5
9	189	42.5
10	192	47.5
11	201	52.5
12	209	57.5
13	210	62.5
14	211	67.5
15	218	72.5
16	238	77.5
17	245	82.5
18	248	87.5
19	253	92.5
20	257	97.5

population from which the data were selected. We can obtain a quantile plot using the following steps:

1. Place a scale on the horizontal axis of a graph covering the interval $(0, 1)$.
2. Place a scale on the vertical axis covering the range of the observed data, y_1 to y_n.
3. Plot $y_{(i)}$ versus $u_i = (i - .5)/n = (2i - 1)/2n$, for $i = 1, \ldots, n$.

Using the Minitab software, we obtain the plot shown in Figure 3.19 for the cholesterol data. Note that, with Minitab, the vertical axis is labeled $Q(u)$ rather than

FIGURE 3.19
Quantile plot of cholesterol data

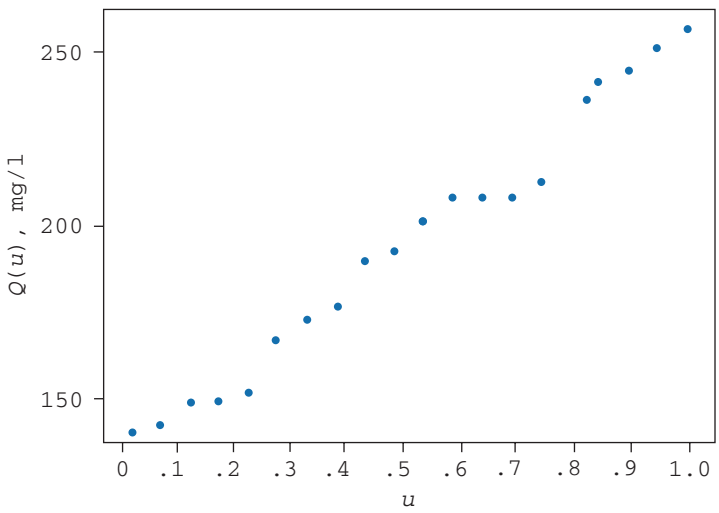
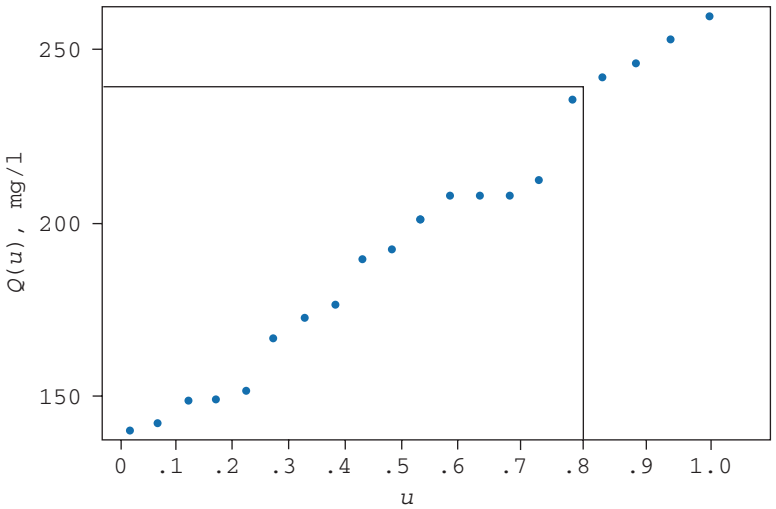


FIGURE 3.20
80th quantile of cholesterol data



$y_{(i)}$. We plot $y_{(i)}$ versus u to obtain a quantile plot. Specific quantiles can be read from the plot.

We can obtain the quantile, $Q(u)$, for any value of u as follows. First, place a smooth curve through the plotted points in the quantile plot, and then read the value off the graph corresponding to the desired value of u.

To illustrate the calculations, suppose we want to determine the 80th percentile for the cholesterol data—that is, the cholesterol level such that 80% of the persons in the population have a cholesterol level less than this value, $Q(.80)$.

Referring to Figure 3.19, locate the point $u = .8$ on the horizontal axis and draw a perpendicular line up to the quantile plot and then a horizontal line over to the vertical axis. The point where this line touches the vertical axis is our estimate of the 80th quantile. (See Figure 3.20.) Roughly 80% of the population has a cholesterol level less than 243. A slightly different definition of the quartiles is given in Section 3.6.

When the data are grouped, the following formula can be used to approximate the percentiles for the original data. Let

P = percentile of interest
L = lower limit of the class interval that includes the percentile of interest
n = total frequency
cf_b = cumulative frequency for all class intervals before the percentile class
f_p = frequency of the class interval that includes the percentile of interest
w = interval width

Then, for example, the 65th percentile for a set of grouped data would be computed using the formula

$$P = L + \frac{w}{f_p}(.65n - cf_b)$$

To determine L, f_p, and cf_b, begin with the lowest interval and find the first interval for which the cumulative relative frequency exceeds .65. This interval would contain the 65th percentile.

EXAMPLE 3.8

Refer to the tick data of Table 3.8. Compute the 90th percentile.

Solution Because the eighth interval is the first interval for which the cumulative relative frequency exceeds .90, we have

$$L = 33.75$$
$$n = 100$$
$$cf_b = 82$$
$$f_{90} = 11$$
$$w = 2.5$$

Thus, the 90th percentile is

$$P_{90} = 33.75 + \frac{2.5}{11}[.9(100) - 82] = 35.57$$

This means that 90% of the cows have 35 or fewer attached ticks and 10% of the cows have 36 or more attached ticks. ■

interquartile range

The second measure of variability, the **interquartile range,** is now defined.

DEFINITION 3.6

The **interquartile range (IQR)** of a set of measurements is defined to be the difference between the upper and lower quartiles; that is,

$$\text{IQR} = \text{75th percentile} - \text{25th percentile}$$

The IQR is displayed in Figure 3.18. The interquartile range, although more sensitive to data pileup about the midpoint than is the range, is still not sufficient for our purposes. In fact, the IQR can be very misleading when the data set is highly concentrated about the median. For example, suppose we have a sample consisting of 10 data values:

$$20, 50, 50, 50, 50, 50, 50, 50, 50, 80$$

The mean, median, lower quartile, and upper quartile would all equal 50. Thus, IQR equals $50 - 50 = 0$. This is very misleading because a measure of variability equal to 0 should indicate that the data consist of *n identical* values, which is not the case in our example. The IQR ignores the extremes in the data set completely. In fact, the IQR measures only the distance needed to cover the middle 50% of the data values and hence totally ignores the spread in the lower and upper 25% of the data. In summary, the IQR does not provide a lot of useful information about the variability of a single set of measurements, but it can be quite useful when comparing the variabilities of two or more data sets. This is especially true when the data sets have some skewness. The IQR will be discussed further in connection with the boxplot (Section 3.6).

In most data sets, we would typically need a minimum of five summary values to provide a minimal description of the data set: smallest value, $y_{(1)}$; lower quartile, $Q(.25)$; median; upper quartile, $Q(.75)$; and largest value, $y_{(n)}$. When the data set has a unimodal, bell-shaped, and symmetric relative frequency histogram, just the

sample mean and a measure of variability, the sample variance, can represent the data set. We will now develop the sample variance.

We seek now a sensitive measure of variability, not only for comparing the variabilities of two sets of measurements but also for interpreting the variability of a single set of measurements. To do this, we work with the **deviation** $y_i - \bar{y}$ of a measurement y_i from the mean $\bar{y}$ of the set of measurements.

deviation

To illustrate, suppose we have five sample measurements $y_1 = 68$, $y_2 = 67$, $y_3 = 66$, $y_4 = 63$, and $y_5 = 61$, which represent the percentages of registered voters in five cities who exercised their right to vote at least once during the past year. These measurements are shown in the dot diagram of Figure 3.21. Each measurement is located by a dot above the horizontal axis of the diagram. We use the sample mean

$$\bar{y} = \frac{\sum_i y_i}{n} = \frac{325}{5} = 65$$

to locate the center of the set, and we construct horizontal lines in Figure 3.21 to represent the deviations of the sample measurements from their mean. The deviations of the measurements are computed by using the formula $y_i - \bar{y}$. The five measurements and their deviations are shown in Figure 3.21.

A data set with very little variability would have most of the measurements located near the center of the distribution. Deviations from the mean for a more variable set of measurements would be relatively large.

Many different measures of variability can be constructed by using the deviation, $y_i - \bar{y}$. A first thought is to use the mean deviation, but this will always equal zero, as it does for our example. A second possibility is to ignore the minus signs and compute the average of the absolute values. However, a more easily interpreted function of the deviations involves the sum of the squared deviations of the measurements from their mean. This measure is called the **variance.**

variance

DEFINITION 3.7

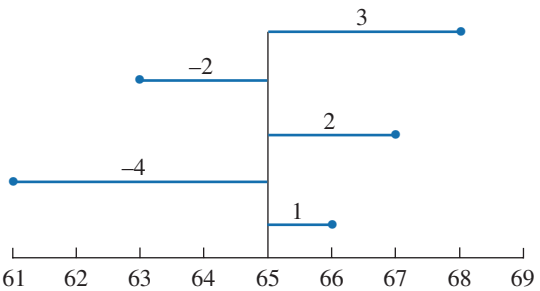
The **variance** of a set of n measurements $y_1, y_2, \ldots, y_n$ with mean $\bar{y}$ is the sum of the squared deviations divided by $n - 1$:

$$\frac{\sum_i (y_i - \bar{y})^2}{n - 1}$$

As with the sample and population means, we have special symbols to denote the sample and population variances. The symbol s^2 represents the sample variance, and the corresponding population variance is denoted by the symbol σ^2.

s^2
σ^2

FIGURE 3.21
Dot diagram of the percentages of registered voters in five cities



The definition for the variance of a set of measurements depends on whether the data are regarded as a sample or population of measurements. The definition we have given here assumes we are working with the sample because the population measurements usually are not available. Many statisticians define the sample variance to be the average of the squared deviations, $\Sigma(y - \bar{y})^2/n$. However, the use of $(n - 1)$ as the denominator of s^2 is not arbitrary. This definition of the sample variance makes it an *unbiased estimator* of the population variance σ^2. This means roughly that if we were to draw a very large number of samples, each of size n, from the population of interest and if we were to compute s^2 for each sample, the average sample variance would equal the population variance σ^2. Had we divided by n in the definition of the sample variance s^2, the average sample variance computed from a large number of samples would be less than the population variance; hence, s^2 would tend to underestimate σ^2.

standard deviation

Another useful measure of variability, the **standard deviation,** involves the square root of the variance. One reason for defining the standard deviation is that it yields a measure of variability having the same units of measurement as the original data, whereas the units for variance are the square of the measurement units.

DEFINITION 3.8

The **standard deviation** of a set of measurements is defined to be the positive square root of the variance.

s
σ

We then have s denoting the sample standard deviation and σ denoting the corresponding population standard deviation.

EXAMPLE 3.9

The time between an electric light stimulus and a bar press to avoid a shock was noted for each of five conditioned rats. Use the given data to compute the sample variance and standard deviation.

Shock avoidance times (seconds): 5, 4, 3, 1, 3

Solution The deviations and the squared deviations are shown in Table 3.11. The sample mean $\bar{y}$ is 3.2.

TABLE 3.11
Shock avoidance data

	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
	5	1.8	3.24
	4	.8	.64
	3	−.2	.04
	1	−2.2	4.84
	3	−.2	.04
Totals	16	0	8.80

Using the total of the squared deviations column, we find the sample variance to be

$$s^2 = \frac{\Sigma_i(y_i - \bar{y})^2}{n - 1} = \frac{8.80}{4} = 2.2 \blacksquare$$

We can make a simple modification of our formula for the sample variance to approximate the sample variance if only grouped data are available. Recall that in approximating the sample mean for grouped data, we let y_i and f_i denote the midpoint and frequency, respectively, for the ith class interval. With this notation, the sample variance for grouped data is $s^2 = \sum_i f_i(y_i - \bar{y})^2/(n-1)$. The sample standard deviation is $\sqrt{s^2}$.

EXAMPLE 3.10

Refer to the tick data from Table 3.9 of Example 3.6. Calculate the sample variance and standard deviation for these data.

Solution From Table 3.9, the sum of the $f_i(y_i - \bar{y})^2$ calculations is 2,704.688. Using this value, we can approximate s^2 and s.

$$s^2 \cong \frac{1}{n-1}\sum_i f_i(y_i - \bar{y})^2 = \frac{1}{99}(2{,}704.688) = 27.32008$$
$$s \cong \sqrt{27.32008} = 5.227$$

If we compute s from the original 100 data values, the value of s (using Minitab) is computed to be 5.212. The values of s computed from the original data and from the grouped data are very close. However, when the frequency table has a small number of classes, the approximation of s from the frequency table values will not generally be as close as in this example. ■

A problem arises with using the standard deviation as a measure of spread in a data set containing a few extreme values. This occurs because the deviations of data values about the mean are squared, resulting in more weight given to the extreme data values. Also, the variance uses the sample/population mean as the central value about which deviations are measured. If a data set contains outliers, a few values that are particularly far away from the mean, either very small or very large, the mean and standard deviation can be overly inflated and hence do not properly represent the center or the spread in the data set. Previously, the median was used in place of the mean to represent the center of the data set when the data set contains outliers. In a similar fashion, an alternative to the standard deviation, the median absolute deviation (MAD) will be defined.

DEFINITION 3.9

The **median absolute deviation** of a set of n measurements $y_1, y_2, \ldots, y_n$ with median $\tilde{y}$ is the median of the absolute deviations of the n measurements about the median:

$$\text{MAD} = \textit{median}\,\{|y_1 - \tilde{y}|, |y_2 - \tilde{y}|, \ldots, |y_n - \tilde{y}|\}/.6745\}$$

EXAMPLE 3.11

Refer to the time between electric light stimulus and a bar press in Example 3.9, and suppose there was a sixth rat in the experiment who had an extremely high tolerance to the shock. This rat had a shock avoidance time of 71 seconds. Compute

the value of the sample standard deviation and MAD for the shock avoidance times for the six values.

Shock avoidance times (seconds): 5, 4, 3, 1, 3, 71

To observe the impact of the extreme value, compare the values of the mean, median, standard deviation, and MAD for the five original shock values to their corresponding values in the new data set.

Solution The deviations, squared deviations, and absolute deviations are given in Table 3.12. The sample mean and median of the six values are, respectively, $\bar{y} = \frac{87}{6} = 14.5$ and $\tilde{y} = \frac{3+4}{2} = 3.5$

TABLE 3.12
Shock avoidance data

Source: *Department of Justice, Crime Reports and the United States, 2000.*

	y_i	$y_i - 14.5$	$(y_i - 14.5)^2$	$y_i - 3.5$	$\lvert y_i - 3.5 \rvert$
	5	−9.5	90.25	1.5	1.5
	4	−10.5	110.25	0.5	0.5
	3	−11.5	132.25	−0.5	0.5
	1	−13.5	182.25	−2.5	2.5
	3	−11.5	132.25	−0.5	0.5
	71	56.5	3,192.25	67.5	67.5
Total	87	0	3,839.50	66.0	73.0

The mean of the six shock times is 14.5 seconds, which is larger than all but one of the six times. The median shock time is 3.5, yielding three shock times less than the median and three shock times greater than the median. Thus, the median is more representative of the center of the data set than is the mean when outliers are present in the data set. The standard deviation is given by

$$s = \sqrt{\frac{\sum_{i=1}^{6}(y_i - 14.5)^2}{6-1}} = \sqrt{\frac{3{,}839.5}{5}} = 27.71$$

MAD is computed as the median of the six absolute deviations about the median divided by 0.6745.

First, compute the median of 0.5, 0.5, 0.5, 1.5, 2.5, and 67.5, which is $(0.5 + 1.5)/2 = 1.0$.

Next, divide the median absolute deviation, 1.0, by 0.6745, yielding MAD = 1.0 /.6745 = 1.48.

The value of the median and MAD from the five shock times in Example 3.9 are 3 and 1.48 compared to 3.5 and 1.48 for the six shock times in the current data set. Thus, the outlier shock time 71 does not have a major impact on the median and MAD as measures of center and spread about the center.

However, the single large shock time greatly inflated the mean and standard deviation, raising the mean from 3.2 to 14.5 seconds and the standard deviation from 1.48 to 27.71 seconds. ■

You may wonder why the median of the absolute deviations is divided by the value 0.6745 in Definition 3.9. In a population having a normal distribution with standard deviation σ, the expected value of the absolute deviation about the median is 0.6745σ. By dividing the median absolute deviation by 0.6745, the

expected value of MAD in a population having a normal distribution is equal to σ. Thus, the values computed for MAD and the sample standard deviation are also the expected values for data randomly selected from populations that have a normal distribution.

We have now discussed several measures of variability, each of which can be used to compare the variabilities of two or more sets of measurements. The standard deviation is particularly appealing for two reasons: (1) We can compare the variabilities of *two or more* sets of data using the standard deviation, and (2) we can also use the results of the rule that follows to interpret the standard deviation of a single set of measurements. This rule applies to data sets with roughly a "mound-shaped" histogram—that is, a histogram that has a single peak, is symmetrical, and tapers off gradually in the tails. Because so many data sets can be classified as mound-shaped, the rule has wide applicability. For this reason, it is called the *Empirical Rule.*

EMPIRICAL RULE

Given a set of n measurements possessing a mound-shaped histogram, then

the interval $\bar{y} \pm s$ contains approximately 68% of the measurements
the interval $\bar{y} \pm 2s$ contains approximately 95% of the measurements
the interval $\bar{y} \pm 3s$ contains approximately 99.7% of the measurements.

EXAMPLE 3.12

The yearly report from a particular stockyard gives the average daily wholesale price per pound for steers as \$.61, with a standard deviation of \$.07. What conclusions can we reach about the daily steer prices for the stockyard? Because the original daily price data are not available, we are not able to provide much further information about the daily steer prices. However, from past experience, it is known that the daily price measurements have a mound-shaped relative frequency histogram. Applying the Empirical Rule, what conclusions can we reach about the distribution of daily steer prices?

Solution Applying the Empirical Rule, the interval

$.61 \pm .07$ or \$.54 to \$.68

contains approximately 68% of the measurements. The interval

$.61 \pm .14$ or \$.47 to \$.75

contains approximately 95% of the measurements. The interval

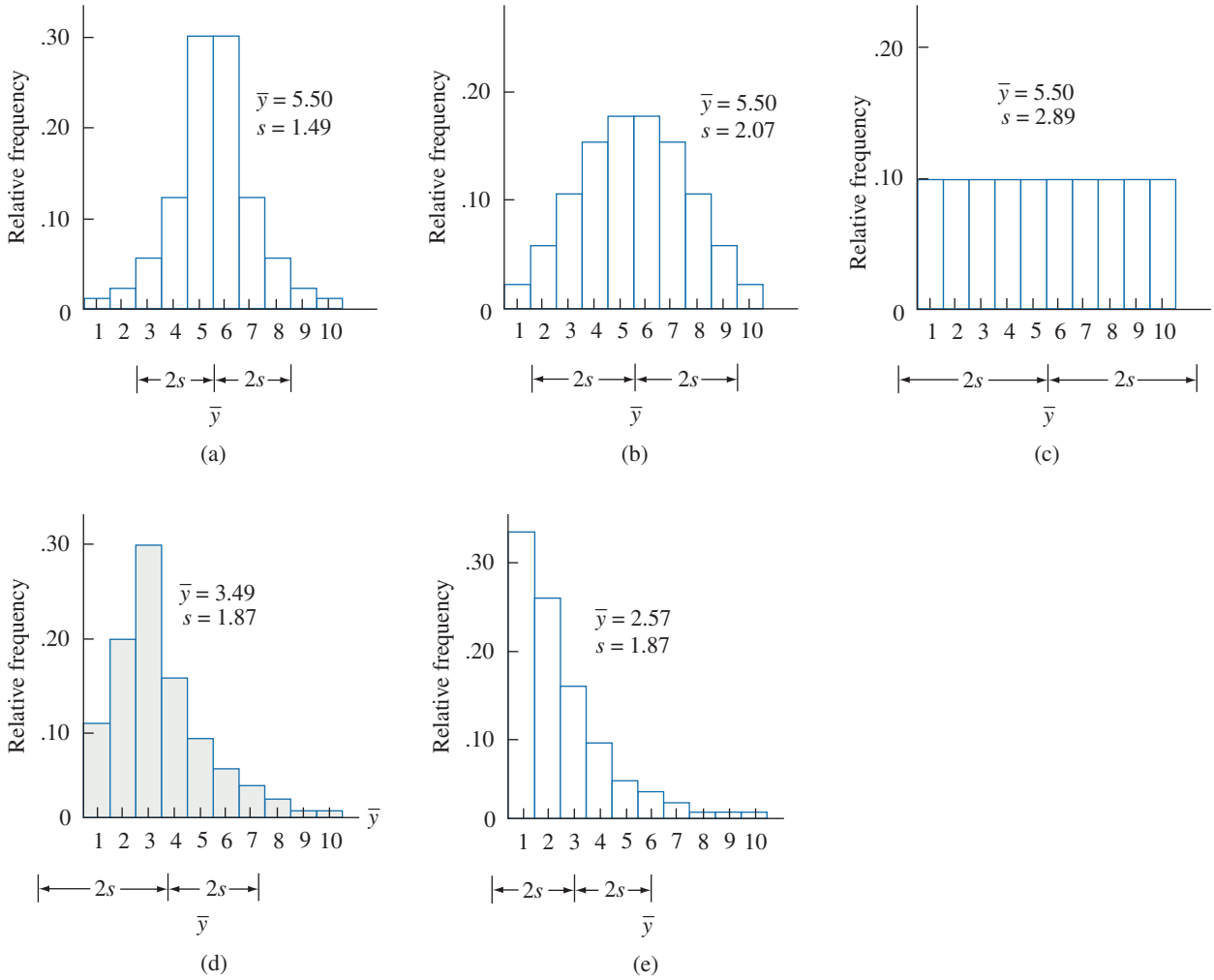
$.61 \pm .21$ or \$.40 to \$.82

contains approximately 99.7% of the measurements. ■

In English, approximately two-thirds of the steers sold for between \$.54 and \$.68 per pound, and 95% sold for between \$.47 and \$.75 per pound, with minimum and maximum prices being approximately \$.40 and \$.82.

To increase our confidence in the Empirical Rule, let us see how well it describes the five frequency distributions of Figure 3.22. We calculated the mean and standard deviation for each of the five data sets (not given), and these are

FIGURE 3.22 A demonstration of the utility of the Empirical Rule



shown next to each frequency distribution. Figure 3.22(a) shows the frequency distribution for measurements made on a variable that can take values $y = 0, 1, 2, \ldots, 10$. The mean and standard deviation $\bar{y} = 5.50$ and $s = 1.49$ for this symmetric, mound-shaped distribution were used to calculate the interval $\bar{y} \pm 2s$, which is marked below the horizontal axis of the graph. We found 94% of the measurements falling in this interval—that is, lying within two standard deviations of the mean. Note that this percentage is very close to the 95% specified in the Empirical Rule. We also calculated the percentage of measurements lying within one standard deviation of the mean. We found this percentage to be 60%, a figure that is not too far from the 68% specified by the Empirical Rule. Consequently, we think the Empirical Rule provides an adequate description for Figure 3.22(a).

Figure 3.22(b) shows another mound-shaped frequency distribution but one that is less peaked than the distribution of Figure 3.22(a). The mean and standard deviation for this distribution, shown to the right of the figure, are 5.50 and 2.07, respectively. The percentages of measurements lying within one and two standard deviations of the mean are 64% and 96%, respectively. Once again, these percentages agree very well with the Empirical Rule.

Now let us look at three other distributions. The distribution in Figure 3.22(c) is perfectly flat, whereas the distributions of Figures 3.22(d) and (e) are nonsymmetric and skewed to the right. The percentages of measurements that lie within two standard deviations of the mean are 100%, 96%, and 95%, respectively, for these three distributions. All these percentages are reasonably close to the 95% specified by the Empirical Rule. The percentages that lie within one standard deviation of the mean (60%, 75%, and 87%, respectively) show some disagreement with the 68% of the Empirical Rule.

To summarize, you can see that the Empirical Rule accurately forecasts the percentage of measurements falling within two standard deviations of the mean for all five distributions of Figure 3.22, even for the distributions that are flat, as in Figure 3.22(c), or highly skewed to the right, as in Figure 3.22(e). The Empirical Rule is less accurate in forecasting the percentage within one standard deviation of the mean, but the forecast, 68%, compares reasonably well for the three distributions that might be called mound-shaped, Figures 3.22(a), (b), and (d).

The results of the Empirical Rule enable us to obtain a quick approximation to the sample standard deviation s. The Empirical Rule states that approximately 95% of the measurements lie in the interval $\bar{y} \pm 2s$. The length of this interval is, therefore, $4s$. Because the range of the measurements is approximately $4s$, we obtain an **approximate value for s** by dividing the range by 4:

approximating *s*

$$\text{approximate value of } s = \frac{\text{range}}{4}$$

Some people might wonder why we did not equate the range to $6s$ because the interval $\bar{y} \pm 3s$ should contain almost all the measurements. This procedure would yield an approximate value for s that is smaller than the one obtained by the preceding procedure. If we are going to make an error (as we are bound to do with any approximation), it is better to overestimate the sample standard deviation so that we are not led to believe there is less variability than may be the case.

EXAMPLE 3.13

The Texas legislature planned on expanding the items on which the state sales tax was imposed. In particular, groceries were previously exempt from sales tax. A consumer advocate argued that lower-income families would be impacted because they spend a much larger percentage of their income on groceries than do middle- and upper-income families. The U.S. Bureau of Labor Statistics publication ***Consumer Expenditures in 2000*** reported that an average family in Texas spent approximately 14% of their family income on groceries. The consumer advocate randomly selected 30 families with income below the poverty level and obtained the following percentages of family incomes allocated to groceries.

26	28	30	37	33	30
29	39	49	31	38	36
33	24	34	40	29	41
40	29	35	44	32	45
35	26	42	36	37	35

For these data, $\sum y_i = 1{,}043$ and $\sum (y_i - \bar{y})^2 = 1{,}069.3667$. Compute the mean, variance, and standard deviation of the percentage of income spent on food. Check your calculation of s.

Solution The sample mean is

$$\bar{y} = \frac{\sum_i y_i}{30} = \frac{1{,}043}{30} = 34.77$$

The corresponding sample variance and standard deviation are

$$s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$$
$$= \frac{1}{29}(1{,}069.3667) = 36.8747$$
$$s = \sqrt{36.8747} = 6.07$$

We can check our calculation of s by using the range approximation. The largest measurement is 49 and the smallest is 24. Hence, an approximate value of s is

$$s \approx \frac{\text{range}}{4} = \frac{49 - 24}{4} = 6.25$$

Note how close the approximation is to our computed value. ■

Although there will not always be the close agreement found in Example 3.13, the range approximation provides a useful and quick check on the calculation of s.

The standard deviation can be deceptive when comparing the amount of variability of different types of populations. A unit of variation in one population might be considered quite small, whereas that same amount of variability in a different population would be considered excessive. For example, suppose we want to compare two production processes that fill containers with products. Process A is filling fertilizer bags, which have a nominal weight of 80 pounds. The process produces bags having a mean weight of 80.6 pounds with a standard deviation of 1.2 pounds. Process B is filling 24-ounce cornflakes boxes, which have a nominal weight of 24 ounces. Process B produces boxes having a mean weight of 24.3 ounces with a standard deviation of 0.4 ounces. Is process A much more variable than process B because 1.2 is three times larger than 0.4? To compare the variability in two considerably different processes or populations, we need to define another measure of variability. The **coefficient of variation** measures the variability in the values in a population relative to the magnitude of the population mean. In a process or population with mean μ and standard deviation σ, the coefficient of variation is defined as

coefficient of variation

$$\text{CV} = \frac{\sigma}{|\mu|}$$

provided $\mu \neq 0$. Thus, the coefficient of variation is the standard deviation of the population or process expressed in units of μ. The two filling processes would have equivalent degrees of variability if the two processes had the same CV. For the fertilizer process, the CV = 1.2/80 = .015. The cornflakes process has CV = 0.4/24 = .017. Hence, the two processes have very similar variability relative to the size of their means. The CV is a unit-free number because the standard deviation and mean are measured using the same units. Hence, the CV is often used as an index of process or population variability. In many applications, the CV is expressed as a percentage: CV = $100(\sigma/|\mu|)\%$. Thus, if a process has a CV of 15%, the standard deviation of the output of the process is 15% of the process mean. Using sampled data from the population, we estimate CV with $100(s/|\bar{y}|)\%$.

3.6 The Boxplot

boxplot

As mentioned earlier in this chapter, a stem-and-leaf plot provides a graphical representation of a set of scores that can be used to examine the shape of the distribution, the range of scores, and where the scores are concentrated. The **boxplot,** which builds on the information displayed in a stem-and-leaf plot, is more concerned with the symmetry of the distribution and incorporates numerical measures of central tendency and location to study the variability of the scores and the concentration of scores in the tails of the distribution.

Before we show how to construct and interpret a boxplot, we need to introduce several new terms that are peculiar to the language of exploratory data analysis (EDA).We are familiar with the definitions for the first, second (median), and third quartiles of a distribution presented earlier in this chapter. The boxplot uses the median and **quartiles** of a distribution.

quartiles

We can now illustrate a *skeletal boxplot* using an example.

EXAMPLE 3.14

A criminologist is studying whether there are wide variations in violent crime rates across the United States. Using Department of Justice data from 2000, the crime rates in 90 cities selected from across the United States were obtained. Use the data given in Table 3.13 to construct a skeletal boxplot to demonstrate the degree of variability in crime rates.

TABLE 3.13
Violent crime rates for 90 standard metropolitan statistical areas selected from around the United States

South	Rate	North	Rate	West	Rate
Albany, GA	498	Allentown, PA	285	Abilene, TX	343
Anderson, SC	676	Battle Creek, MI	490	Albuquerque, NM	946
Anniston, AL	344	Benton Harbor, MI	528	Anchorage, AK	584
Athens, GA	368	Bridgeport, CT	427	Bakersfield, CA	494
Augusta, GA	772	Buffalo, NY	413	Brownsville, TX	463
Baton Rouge, LA	497	Canton, OH	220	Denver, CO	357
Charleston, SC	415	Cincinnati, OH	163	Fresno, CA	761
Charlottesville, VA	925	Cleveland, OH	428	Galveston, TX	717
Chattanooga, TN	555	Columbus, OH	625	Houston, TX	1094
Columbus, GA	260	Dayton, OH	339	Kansas City, MO	637
Dothan, AL	528	Des Moines, IA	211	Lawton, OK	692
Florence, SC	649	Dubuque, IA	451	Lubbock, TX	522
Fort Smith, AR	571	Gary, IN	358	Merced, CA	397
Gadsden, AL	470	Grand Rapids, MI	660	Modesto, CA	521
Greensboro, NC	897	Janesville, WI	330	Oklahoma City, OK	610
Hickery, NC	973	Kalamazoo, MI	145	Reno, NV	477
Knoxville, TN	486	Lima, OH	326	Sacramento, CA	453
Lake Charles, LA	447	Madison, WI	403	St. Louis, MO	798
Little Rock, AR	689	Milwaukee, WI	523	Salinas, CA	646
Macon, GA	754	Minneapolis, MN	312	San Diego, CA	645
Monroe, LA	465	Nassau, NY	576	Santa Ana, CA	549
Nashville, TN	496	New Britain, CT	261	Seattle, WA	568
Norfolk, VA	871	Philadelphia, PA	221	Sioux City, IA	465
Raleigh, NC	1064	Pittsburgh, PA	754	Stockton, CA	350

South	Rate	North	Rate	West	Rate
Richmond, VA	579	Portland, ME	140	Tacoma, WA	574
Savannah, GA	792	Racine, WI	418	Tucson, AZ	944
Shreveport, LA	367	Reading, PA	657	Victoria, TX	426
Washington, DC	998	Saginaw, MI	564	Waco, TX	477
Wilmington, DE	773	Syracuse, NY	405	Wichita Falls, TX	354
Wilmington, NC	887	Worcester, MA	872	Yakima, WA	264

Note: Rates represent the number of violent crimes (murder, forcible rape, robbery, and aggravated assault) per 100,000 inhabitants, rounded to the nearest whole number.

Source: Department of Justice, Crime Reports and the United States, 2000.

Solution The data were summarized using a stem-and-leaf plot as depicted in Figure 3.23. Use this plot to construct a skeletal boxplot.

FIGURE 3.23
Stem-and-leaf plot of crime data

```
 1  40  45
 1  63
 2  11  20  21
 2  60  61  64  85
 3  12  26  30  39  43  44
 3  50  54  57  58  67  68  97
 4  03  05  13  15  18  26  27  28  47
 4  51  53  63  65  65  70  77  77  86  90  94  96  97  98
 5  21  22  23  28  28  49
 5  55  64  68  71  74  76  79  84
 6  10  25  37  45  46  49
 6  57  60  76  89  92
 7  17
 7  54  54  61  72  73  92  98
 8
 8  71  72  87  97
 9  25  44  46
 9  73  98
10
10  64  94
```

When the scores are ordered from lowest to highest, the median is computed by averaging the 45th and 46th scores. For these data, the 45th score (counting from the lowest to the highest in Figure 3.23) is 497 and the 46th is 498; hence, the median is

$$M = \frac{497 + 498}{2} = 497.5$$

To find the lower and upper quartiles for this distribution of scores, we need to determine the 25th and 75th percentiles. We can use the method given on page 94 to compute $Q(.25)$ and $Q(.75)$. A quick method that yields essentially the same values for the two quartiles consists of the following steps:

1. Order the data from smallest to largest value.
2. Divide the ordered data set into two data sets using the median as the dividing value.

3. Let the lower quartile be the median of the set of values consisting of the smaller values.
4. Let the upper quartile be the median of the set of values consisting of the larger values.

In the example, the data set has 90 values. Thus, we create two data sets, one containing the $90/2 = 45$ smallest values, and the other containing the 45 largest values. The lower quartile is the $(45 + 1)/2 = 23$rd smallest value, and the upper quartile is the 23rd value counting from the largest value in the data set. The 23rd-lowest score and 23rd-highest score are 397 and 660.

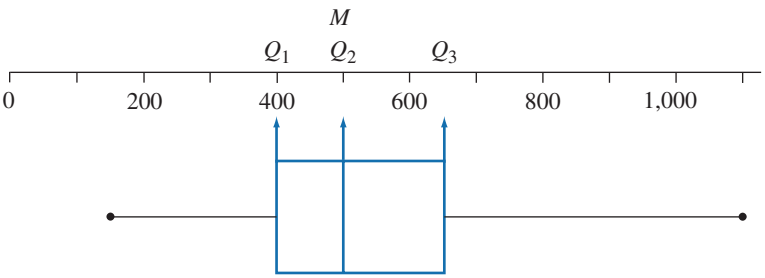
$$\text{lower quartile, } Q_1 = 397$$
$$\text{upper quartile, } Q_3 = 660$$

skeletal boxplot

These three descriptive measures and the smallest and largest values in a data set are used to construct a skeletal boxplot (see Figure 3.24). The **skeletal boxplot** is constructed by drawing a box between the lower and upper quartiles with a solid line drawn across the box to locate the median. A straight line is then drawn connecting the box to the largest value; a second line is drawn from the box to the smallest value. These straight lines are sometimes called whiskers, and the entire graph is called a skeletal **box-and-whiskers plot.**

box-and-whiskers plot

FIGURE 3.24
Skeletal boxplot for the data of Figure 3.23



With a quick glance at a skeletal boxplot, it is easy to obtain an impression about the following aspects of the data:

1. The lower and upper quartiles, Q_1 and Q_3
2. The interquartile range (IQR), the distance between the lower and upper quartiles
3. The most extreme (lowest and highest) values
4. The symmetry or asymmetry of the distribution of scores

If we were presented with Figure 3.24 without having seen the original data, we would have observed that

$$Q_1 \approx 400$$
$$Q_3 \approx 675$$
$$\text{IQR} \approx 675 - 400 = 275$$
$$M \approx 500$$
$$\text{most extreme values: } \approx 150 \text{ and } \approx 1{,}100$$

Also, because the median is closer to the lower quartile than the upper quartile and because the upper whisker is a little longer than the lower whisker, the distribution is slightly nonsymmetrical. To see that this conclusion is true, construct a frequency histogram for these data.

The skeletal boxplot can be expanded to include more information about extreme values in the tails of the distribution. To do so, we need the following additional quantities:

lower inner fence: $Q_1 - 1.5(\text{IQR})$
upper inner fence: $Q_3 + 1.5(\text{IQR})$
lower outer fence: $Q_1 - 3(\text{IQR})$
upper outer fence: $Q_3 + 3(\text{IQR})$

Any data value beyond an inner fence on either side is called a *mild outlier,* and any data value beyond an outer fence on either side is called an *extreme outlier.* The smallest and largest data values that are *not* outliers are called the *lower adjacent value* and *upper adjacent value*, respectively.

EXAMPLE 3.15

Compute the inner and outer fences for the data of Example 3.14. Identify any mild and extreme outliers.

Solution For these data, we found the lower and upper quartiles to be 397 and 660, respectively; IQR = 660 − 397 = 263. Then

lower inner fence = 397 − 1.5(263) = 2.5
upper inner fence = 660 + 1.5(263) = 1,054.5
lower outer fence = 397 − 3(263) = −392
upper outer fence = 660 + 3(263) = 1,449

Also, from the stem-and-leaf plot, we can determine that the lower and upper adjacent values are 140 and 998. There are two mild outliers, 1,064 and 1,094, because both values fall between the upper inner fence, 1,054.5, and upper outer fence, 1,449. ■

We now have all the quantities necessary for constructing a boxplot, sometimes refered to as a modified boxplot.

Steps in Constructing a Boxplot

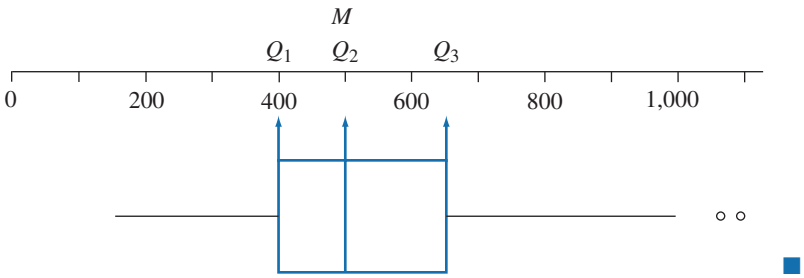
1. As with a skeletal boxplot, mark off a box from the lower quartile to the upper quartile.
2. Draw a solid line across the box to locate the median.
3. Draw a line from each quartile to its adjacent value.
4. Mark each mild outlier with an open circle, ○.
5. Mark each extreme outlier with a closed circle, ●.

EXAMPLE 3.16

Construct a boxplot for the data of Example 3.13.

Solution The boxplot is shown in Figure 3.25.

FIGURE 3.25
Boxplot for the data of Example 3.13

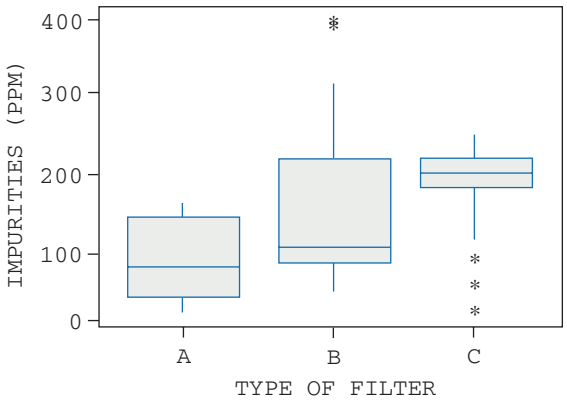


What information can be drawn from a boxplot? First, the center of the distribution of scores is indicated by the median line (Q_2) in the boxplot. Second, a measure of the variability of the scores is given by the interquartile range, the length of the box. Recall that the box is constructed between the lower and upper quartiles, so it contains the middle 50% of the scores in the distribution, with 25% on either side of the median line inside the box. Third, by examining the relative position of the median line, we can gauge the symmetry of the middle 50% of the scores. For example, if the median line is closer to the lower quartile than the upper, there is a greater concentration of scores on the lower side of the median within the box than on the upper side; a symmetric distribution of scores would have the median line located in the center of the box. Fourth, additional information about skewness is obtained from the lengths of the whiskers; the longer one whisker is relative to the other one, the more skewness there is in the tail with the longer whisker. Fifth, a general assessment can be made about the presence of outliers by examining the number of scores classified as mild outliers and the number classified as extreme outliers.

Boxplots provide a powerful graphical technique for comparing samples from several different treatments or populations. We will illustrate these concepts using the following example. Several new filtration systems have been proposed for use in small city water systems. The three systems under consideration have very similar initial and operating costs, and will be compared on the basis of the amount of impurities remaining in the water after it passes through the system. After careful assessment, it is determined that monitoring 20 days of operation will provide sufficient information to determine any significant differences among the three systems. Water samples are collected on a hourly basis. The amount of impurities, in ppm, remaining in the water after the water passes through the filter is recorded. The average daily values for the three systems are plotted using a side-by-side boxplot, as presented in Figure 3.26.

An examination of the boxplots in Figure 3.26 reveals the shapes of the relative frequency histograms for the three types of filters based on their boxplots. Filter A has a symmetric distribution, filter B is skewed to the right, and filter C is skewed to the left. Filters A and B have nearly equal medians. However, filter B is much more variable than both filters A and C. Filter C has a larger median than both filters A and B but smaller variability than A with the exception of the two very small values obtained using filter C. The mild outliers obtained by filters B and C, identified by *, would be examined to make sure that they are valid measurements. Note that the software package, Minitab, used to produce the graph, uses the symbol * in place of the open circle ○ to designate a mild outlier. These measurements could be either recording errors or operational errors. They must be carefully checked because they have such a large influence on the summary

FIGURE 3.26
Removing impurities using three filter types



statistics. Filter A would produce a more consistent filtration than filter B. Filter A generally filters the water more thoroughly than filter C. We will introduce statistical techniques in Chapter 8 that will provide us with ways to differentiate among the three filter types.

3.7 Summarizing Data from More Than One Variable: Graphs and Correlation

In the previous sections, we've discussed graphical methods and numerical descriptive methods for summarizing data from a single variable. Frequently, more than one variable is being studied at the same time, and we might be interested in summarizing the data on each variable separately and also in studying relations among the variables. For example, we might be interested in the prime interest rate and in the Consumer Price Index, as well as in the relation between the two. In this section, we'll discuss a few techniques for summarizing data from two (or more) variables. Material in this section will provide a brief preview of and introduction to contingency tables (Chapter 10), analysis of variance (Chapters 8 and 14–18), and regression (Chapters 11, 12, and 13).

contingency table

Consider first the problem of summarizing data from two qualitative variables. Cross-tabulations can be constructed to form a **contingency table.** The rows of the table identify the categories of one variable, and the columns identify the categories of the other variable. The entries in the table are the number of times each value of one variable occurs with each possible value of the other. For example, episodic or "binge" drinking—the consumption of large quantities of alcohol at a single session resulting in intoxication—among eighteen- to twenty-four-year-olds can have a wide range of adverse effects—medical, personal, and social. A survey was conducted on 917 eighteen- to twenty-four-year-olds by the ***Institute of Alcohol Studies***. Each individual surveyed was asked questions about his or her alcohol consumption in the prior 6 months. The criminal background of the individuals was also obtained from a police data base. The results of the survey are displayed in Table 3.14. From this table, it is observed that 114 of binge drinkers were involved in violent crimes, whereas 27 occasional drinkers and 7 nondrinkers were involved in violent crimes.

One method for examining the relationships between variables in a contingency table is a percentage comparison based on row totals, column totals, or the overall total. If we calculate percentages within each column, we can compare

TABLE 3.14
Data from a survey of drinking behavior of eighteen- to twenty-four-year-old youths

	Level of Drinking			
Criminal Offense	**Binge/Regular Drinker**	**Occasional Drinker**	**Never Drinks**	**Total**
Violent Crime	114	27	7	148
Theft/Property Damage	53	27	7	87
Other Criminal Offenses	138	53	15	206
No Criminal Offenses	50	274	152	476
Total	355	381	181	917

Source: ***Institute of Alcohol Studies.***

TABLE 3.15
Comparing the distribution of criminal activity for each level of alcohol consumption

	Level of Drinking		
Criminal Offense	**Binge/Regular Drinker**	**Occasional Drinker**	**Never Drinks**
Violent Crime	32.1%	7.1%	3.9%
Theft/Property Damage	14.9%	7.1%	3.9%
Other Criminal Offenses	38.9%	13.9%	8.2%
No Criminal Offenses	14.1%	71.9%	84.0%
Total	100% (n = 355)	100% (n = 381)	100% (n = 181)

the distribution of criminal activity within each level of drinking. A percentage comparison based on column totals is shown in Table 3.15.

For all three types of criminal activities, the binge/regular drinkers had more than double the level of activity of the occassional or nondrinkers. For binge/regular drinkers, 32.1% had committed a violent crime, whereas only 7.1% of occasional drinkers and 3.9% of nondrinkers had committed a violent crime. This pattern is repeated across the other two levels of criminal activity. In fact, 85.9% of binge/regular drinkers had committed some form of criminal violation. The level of criminal activity among occasional drinkers was 28.1% and only 16% for nondrinkers. In Chapter 10, we will use statistical methods to explore further relations between two (or more) qualitative variables.

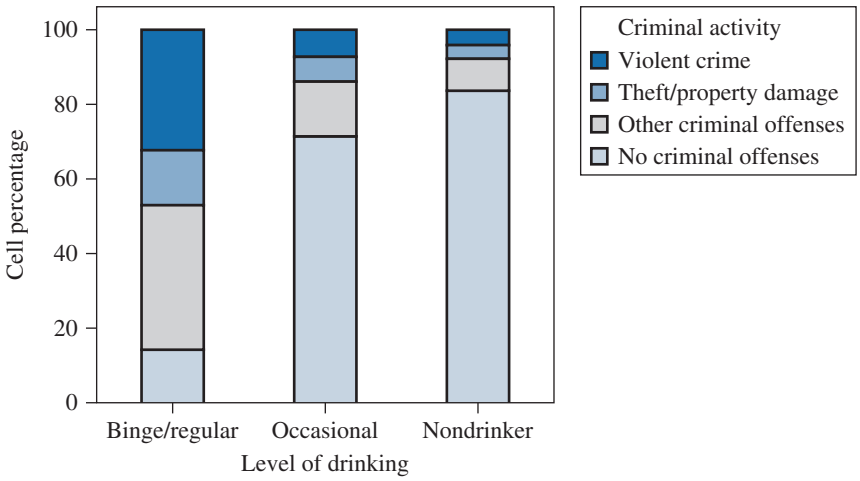
stacked bar graph

An extension of the bar graph provides a convenient method for displaying data from a pair of qualitative variables. Figure 3.27 is a **stacked bar graph,** which displays the data in Table 3.15.

The graph represents the distribution of criminal activity for three levels of alcohol consumption by young adults. This type of information is useful in making youths aware of the dangers involved in the consumption of large amounts of alcohol. While the heaviest drinkers are at the greatest risk of committing a criminal offense, the risk of increased criminal behavior is also present for occasional drinkers when compared to those youths who are nondrinkers. This type of data may lead to programs that advocate prevention policies and assistance from the beer/alcohol manufacturers by including messages about appropriate consumption in their advertising.

A second extension of the bar graph provides a convenient method for displaying the relationship between a single quantitative and a single qualitative variable. A food scientist is studying the effects of combining different types of

FIGURE 3.27
Chart of cell percentages for level of drinking versus criminal activity



fats with different surfactants on the specific volume of baked bread loaves. The experiment is designed with three levels of surfactant and three levels of fat, a 3×3 factorial experiment with varying number of loaves baked from each of the nine treatments. She bakes bread from dough mixed from the nine different combinations of the types of fat and types of surfactants and then measures the specific volume of the bread. The data and summary statistics are displayed in Table 3.16.

cluster bar graph

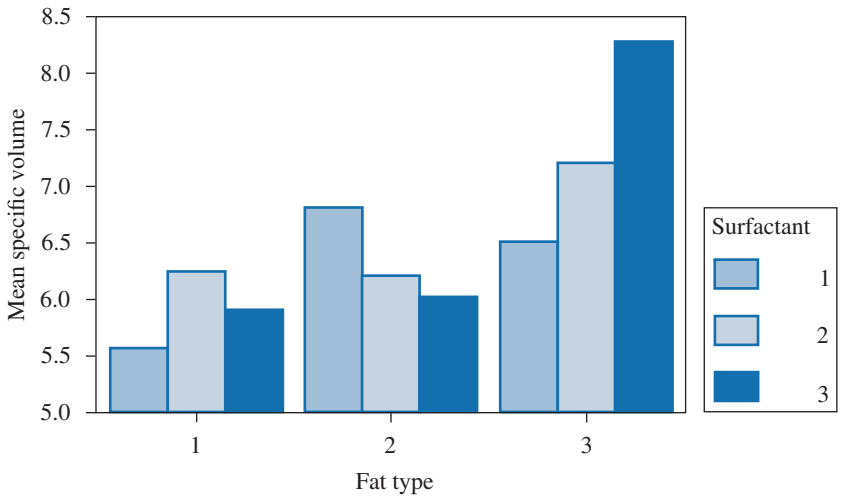
In this experiment, the scientist wants to make inferences from the results of the experiment for the commercial production process. Figure 3.28 is a **cluster bar graph** from the baking experiment. This type of graph allows the experimenter to examine the simultaneous effects of two factors, type of fat and type of surfactant, on the specific volume of the bread. Thus, the researcher can examine the differences in the specific volumes of the nine different ways in which the bread was formulated. A quantitative assessment of the effects of fat type and surfactant type on the mean specific volume will be addressed in Chapter 15.

We can also construct data plots to summarize the relation between two quantitative variables. Consider the following example. A manager of a small

TABLE 3.16
Descriptive statistics with the dependent variable, specific volume

Fat	Surfactant	Mean	Standard Deviation	*N*
1	1	5.567	1.206	3
	2	6.200	.794	3
	3	5.900	.458	3
	Total	5.889	.805	9
2	1	6.800	.794	3
	2	6.200	.849	2
	3	6.000	.606	4
	Total	6.311	.725	9
3	1	6.500	.849	2
	2	7.200	.668	4
	3	8.300	1.131	2
	Total	7.300	.975	8
Total	1	6.263	1.023	8
	2	6.644	.832	9
	3	6.478	1.191	9
	Total	6.469	.997	26

FIGURE 3.28
Specific volumes from baking experiment



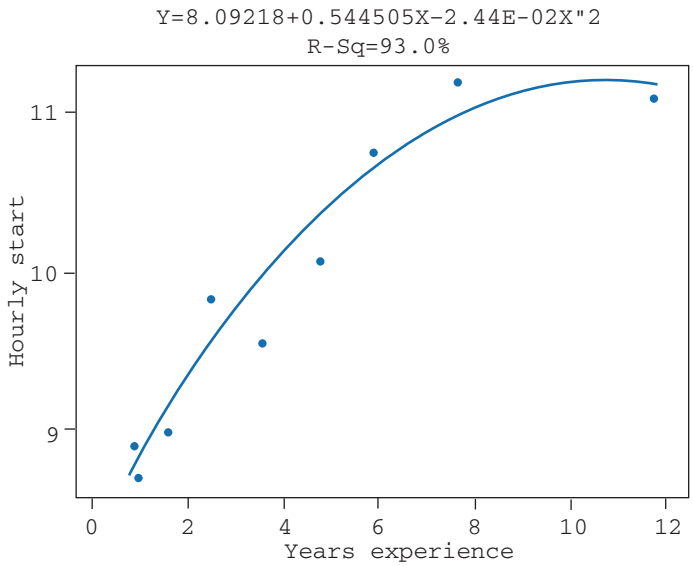
machine shop examined the starting hourly wage y offered to machinists with x years of experience. The data are shown here:

y (dollars)	8.90	8.70	9.10	9.00	9.79	9.45	10.00	10.65	11.10	11.05
x (years)	1.25	1.50	2.00	2.00	2.75	4.00	5.00	6.00	8.00	12.00

scatterplot

Is there a relationship between hourly wage offered and years of experience? One way to summarize these data is to use a **scatterplot,** as shown in Figure 3.29. Each point on the plot represents a machinist with a particular starting wage and years of experience. The smooth curve fitted to the data points, called the *least squares line,* represents a summarization of the relationship between y and x. This line allows the prediction of hourly starting wages for a machinist having years of experience not represented in the data set. How this curve is obtained will be discussed in Chapters 11 and 12. In general, the fitted curve indicates that, as the years of experience x increase, the hourly starting wage increases to a point and then levels

FIGURE 3.29
Scatterplot of starting hourly wage and years of experience



off. The basic idea of relating several quantitative variables is discussed in the chapters on regression (Chapters 11–13).

Using a scatterplot, the general shape and direction of the relationship between two quantitative variables can be displayed. In many instances, the relationship can be summarized by fitting a straight line through the plotted points. Thus, the strength of the relationship can be described in the following manner. There is a strong relationship if the plotted points are positioned close to the line and a weak relationship if the points are widely scattered about the line. It is fairly difficult to "eyeball" the strength using a scatterplot. In particular, if we wanted to compare two different scatterplots, a numerical measure of the strength of the relationship would be advantagous. The following example will illustrate the difficulty of using scatterplots to compare the strength of the relationship between two quantitative variables.

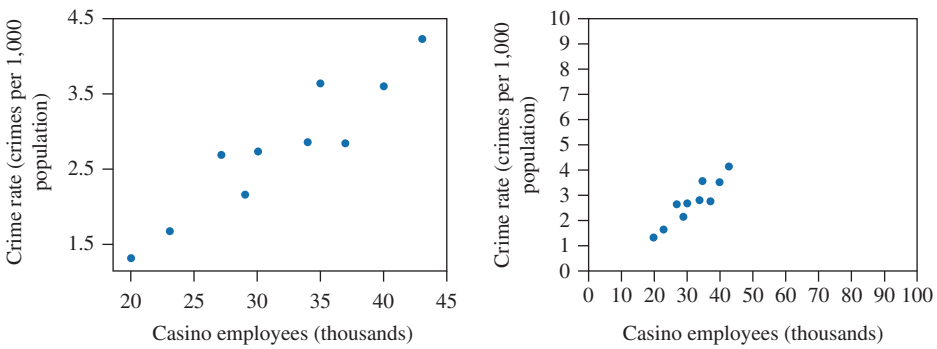
Several major cities in the United States are now considering allowing gambling casinos to operate under their jurisdiction. A major argument in opposition to casino gambling is the perception that there will be a subsequent increase in the crime rate. Data were collected over a 10-year period in a major city where casino gambling had been legalized. The results are listed in Table 3.17 and plotted in Figure 3.30. The two scatterplots are depicting exactly the same data, but the scales of the plots differ considerably. The results appear to show a stronger relationship in one scatterplot that in the other.

Because of the difficulty of determining the strength of the relationship between two quantitative variables by visually examining a scatterplot, a numerical measure of the strength of the relationship will be defined as a supplement to a

TABLE 3.17
Crime rate as a function of number of casino employees

Year	Number of Casino Employees x (thousands)	Crime Rate y (number of crimes per 1,000 population)
1994	20	1.32
1995	23	1.67
1996	29	2.17
1997	27	2.70
1998	30	2.75
1999	34	2.87
2000	35	3.65
2001	37	2.86
2002	40	3.61
2003	43	4.25

FIGURE 3.30
Crime rate as a function of number of casino employees



graphical display. The *correlation coefficient* was first introduced by Francis Galton in 1888. He applied the correlation coefficient to study the relationship between the forearm length and height of particular groups of people.

DEFINITION 3.10

The **correlation coefficient** measures the strength of the linear relationship between two quantitative variables. The correlation coefficient is usually denoted as r.

Suppose we have data on variables x and y collected from n individuals or objects, with means and standard deviations of the variables given as $\bar{x}$ and s_x for the x-variable and $\bar{y}$ and s_y for the y-variable. The correlation r between x and y is computed as

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right] / s_x s_y$$

In computing the correlation coefficient, the variables x and y are standardized to be unit-free variables. The standardized x-variable for the ith individual, $\left(\frac{x_i - \bar{x}}{s_x}\right)$, measures how many standard deviations x_i is above or below the x-mean. Thus, the correlation coefficient, r, is a unit-free measure of the strength of the linear relationship between the quantitative variables, x and y.

EXAMPLE 3.16

For the data in Table 3.17, compute the value of the correlation coefficient.

Solution The computation of r can be performed by any of the statistical software packages or by Excel. The calculations required to obtain the value of r for the data in Table 3.17 are given in Table 3.18, with $\bar{x} = 31.80$ and $\bar{y} = 2.785$. The first row is computed as

$$x - \bar{x} = 20 - 31.8 = -11.8, \quad y - \bar{y} = 1.32 - 2.785 = -1.465,$$
$$(x - \bar{x})(y - \bar{y}) = (-11.8)(-1.465) = 17.287,$$
$$(x - \bar{x})^2 = (-11.8)^2 = 139.24, \quad (y - \bar{y})^2 = (-1.465)^2 = 2.14623$$

TABLE 3.18
Data and calculations for computing r

	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
	20	1.32	−11.8	−1.465	17.287	139.24	2.14623
	23	1.67	−8.8	−1.115	9.812	77.44	1.24323
	29	2.17	−2.8	−0.615	1.722	7.84	0.37823
	27	2.70	−4.8	−0.085	0.408	23.04	0.00722
	30	2.75	−1.8	−0.035	0.063	3.24	0.00123
	34	2.87	2.2	0.085	0.187	4.84	0.00722
	35	3.65	3.2	0.865	2.768	10.24	0.74822
	37	2.86	5.2	0.075	0.390	27.04	0.00562
	40	3.61	8.2	0.825	6.765	67.24	0.68062
	43	4.25	11.2	1.465	16.408	125.44	2.14622
Total	318	27.85	0	0	55.810	485.60	7.3641
Mean	31.80	2.785					

A form of r that is somewhat more direct in its calculation is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{55.810}{\sqrt{(485.6)(7.3641)}} = .933$$

The above calculations depict a positive correlation between the number of casino employees and the crime rate. However, this result does not prove that an increase in the number of casino workers *causes* an increase in the crime rate. There may be many other associated factors involved in the increase of the crime rate.

Generally, the correlation coefficient, r, is a positive number if y tends to increase as x increases; r is negative if y tends to decrease as x increases; and r is nearly zero if there is either no relation between changes in x and changes in y or a nonlinear relation between x and y such that the patterns of increase and decrease in y (as x increases) cancel each other.

Some properties of r that assist us in the interpretation of the relationship between two variables include the following:

1. A positive value for r indicates a positive association between the two variables, and a negative value for r indicates a negative association between the two variables.
2. The value of r is a number between -1 and $+1$. When the value of r is very close to ± 1, the points in the scatterplot will lie close to a straight line.
3. Because the two variables are standardized in the calculation of r, the value of r does not change if we alter the units of x or y. The same value of r will be obtained no matter what units are used for x and y. Correlation is a unit-free measure of association.
4. Correlation measures the degree of the straight-line relationship between two variables. The correlation coefficient does *not* describe the closeness of the points (x, y) to a curved relationship, no matter how strong the relationship.

What values of r indicate a "strong" relationship between y and x? Figure 3.31 displays 15 scatterplots obtained by randomly selecting 1,000 pairs (x_i, y_i) from 15 populations having bivariate normal distributions with correlations ranging from $-.99$ to .99. We can observe that unless $|r|$ is greater than .6, there is very little trend in the scatterplot.

Finally, we can construct data plots for summarizing the relations among several quantitative variables. Consider the following example. Thall and Vail (1990) described a study to evaluate the effectiveness of the anti-epileptic drug progabide as an adjuvant to standard chemotherapy. A group of 59 epileptics was selected to be used in the clinical trial. The patients suffering from simple or complex partial seizures were randomly assigned to receive either the anti-epileptic drug progabide or a placebo. At each of four successive postrandomization clinic visits, the number of seizures occurring over the previous 2 weeks was reported. The measured variables were y_i $(i = 1, 2, 3, 4)$, the seizure counts recorded at the four clinic visits; Trt (x_1), where 0 is the placebo and 1 is progabide; Base (x_2), the baseline seizure rate; and Age (x_3), the patient's age in years. The data and summary statistics are given in Tables 3.19 and 3.20.

side-by-side boxplots

The first plots are **side-by-side boxplots** that compare the base number of seizures and the age of the treated patients to those of the patients assigned to the placebo. These plots provide a visual assessment of whether the treated patients

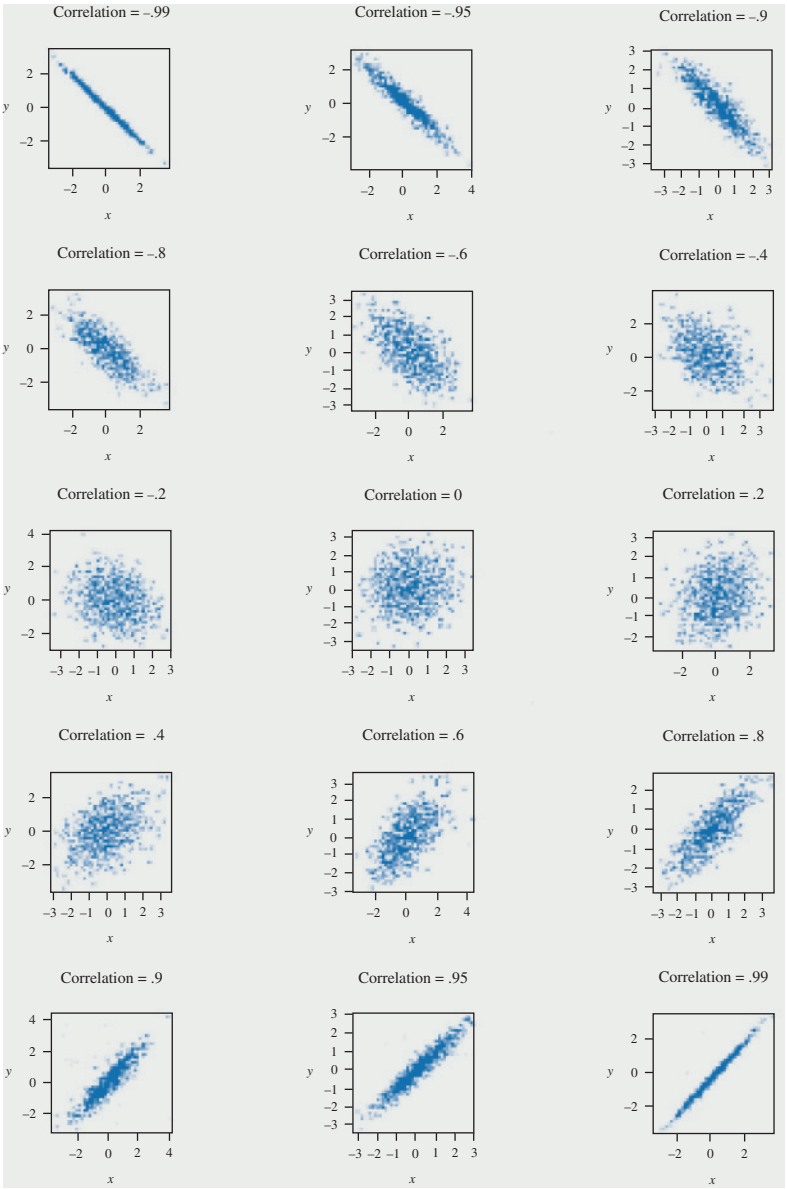


FIGURE 3.31
Scatterplots showing various values for *r*

and placebo patients had similar distributions of age and base seizure counts prior to the start of the clinical trials. An examination of Figure 3.32(a) reveals that the seizure patterns prior to the beginning of the clinical trials are similar for the two groups of patients. There is a single patient with a base seizure count greater than 100 in both groups. The base seizure count for the placebo group is somewhat more variable than that for the treated group—its box is wider than the box for the treated group. The descriptive statistics table contradicts this observation. The sample standard deviation is 26.10 for the placebo group and 27.98 for the treated group. This seemingly inconsistent result occurs due to the large base count for a single patient in the treated group. The median number of base seizures is higher for the treated group than for the placebo group. The means are nearly identical for the two groups. The means are in greater agreement than are the medians due

TABLE 3.19
Data for epilepsy study: successive 2-week seizure counts for 59 epileptics; covariates are adjuvant treatment (0 = placebo, 1 = progabide), 8-week baseline seizure counts, and age (in years)

ID	y_1	y_2	y_3	y_4	Trt	Base	Age
104	5	3	3	3	0	11	31
106	3	5	3	3	0	11	30
107	2	4	0	5	0	6	25
114	4	4	1	4	0	8	36
116	7	18	9	21	0	66	22
118	5	2	8	7	0	27	29
123	6	4	0	2	0	12	31
126	40	20	23	12	0	52	42
130	5	6	6	5	0	23	37
135	14	13	6	0	0	10	28
141	26	12	6	22	0	52	36
145	12	6	8	4	0	33	24
201	4	4	6	2	0	18	23
202	7	9	12	14	0	42	36
205	16	24	10	9	0	87	26
206	11	0	0	5	0	50	26
210	0	0	3	3	0	18	28
213	37	29	28	29	0	111	31
215	3	5	2	5	0	18	32
217	3	0	6	7	0	20	21
219	3	4	3	4	0	12	29
220	3	4	3	4	0	9	21
222	2	3	3	5	0	17	32
226	8	12	2	8	0	28	25
227	18	24	76	25	0	55	30
230	2	1	2	1	0	9	40
234	3	1	4	2	0	10	19
238	13	15	13	12	0	47	22
101	11	14	9	8	1	76	18
102	8	7	9	4	1	38	32
103	0	4	3	0	1	19	20
108	3	6	1	3	1	10	30
110	2	6	7	4	1	19	18
111	4	3	1	3	1	24	24
112	22	17	19	16	1	31	30
113	5	4	7	4	1	14	35
117	2	4	0	4	1	11	27
121	3	7	7	7	1	67	20
122	4	18	2	5	1	41	22
124	2	1	1	0	1	7	28
128	0	2	4	0	1	22	23
129	5	4	0	3	1	13	40
137	11	14	25	15	1	46	33
139	10	5	3	8	1	36	21
143	19	7	6	7	1	38	35
147	1	1	2	3	1	7	25
203	6	10	8	8	1	36	26
204	2	1	0	0	1	11	25
207	102	65	72	63	1	151	22
208	4	3	2	4	1	22	32
209	8	6	5	7	1	41	25
211	1	3	1	5	1	32	35
214	18	11	28	13	1	56	21
218	6	3	4	0	1	24	41
221	3	5	4	3	1	16	32
225	1	23	19	8	1	22	26
228	2	3	0	1	1	25	21
232	0	0	0	0	1	13	36
236	1	4	3	2	1	12	37

TABLE 3.20
Descriptive statistics: Minitab output for epilepsy example (worksheet size: 100,000 cells)

```
0=PLACEBO
1=TREATED

Variable    TREATMENT       N     Mean   Median  Tr Mean   StDev  SE Mean
Y1          0              28     9.36     5.00     8.54   10.14     1.92
            1              31     8.58     4.00     5.26   18.24     3.28

Y2          0              28     8.29     4.50     7.81    8.16     1.54
            1              31     8.42     5.00     6.37   11.86     2.13

Y3          0              28     8.79     5.00     6.54   14.67     2.77
            1              31     8.13     4.00     5.63   13.89     2.50

Y4          0              28     7.96     5.00     7.46    7.63     1.44
            1              31     6.71     4.00     4.78   11.26     2.02

BASE        0              28    30.79    19.00    28.65   26.10     4.93
            1              31    31.61    24.00    27.37   27.98     5.03

AGE         0              28    29.00    29.00    28.88    6.00     1.13
            1              31    27.74    26.00    27.52    6.60     1.19

Variable    TREATMENT     Min      Max       Q1       Q3
Y1          0            0.00    40.00     3.00    12.75
            1            0.00   102.00     2.00     8.00

Y2          0            0.00    29.00     3.00    12.75
            1            0.00    65.00     3.00    10.00

Y3          0            0.00    76.00     2.25     8.75
            1            0.00    72.00     1.00     8.00

Y4          0            0.00    29.00     3.00    11.25
            1            0.00    63.00     2.00     8.00

BASE        0            6.00   111.00    11.00    49.25
            1            7.00   151.00    13.00    38.00

AGE         0           19.00    42.00    24.25    32.00
            1           18.00    41.00    22.00    33.00
```

FIGURE 3.32(a)
Boxplot of base by treatment

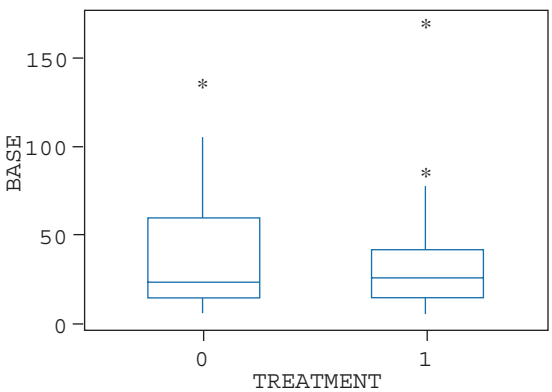
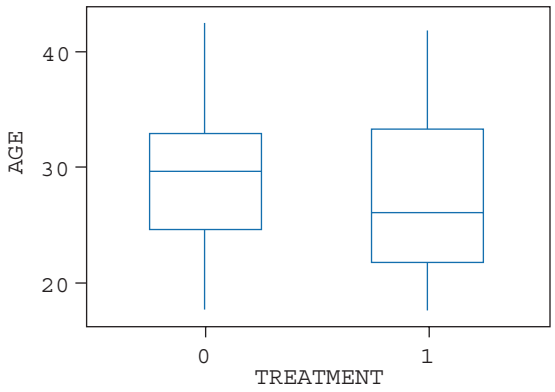


FIGURE 3.32(b)
Boxplot of age by treatment



to the skewed-to-the-right distribution of the middle 50% of the data for the placebo group, whereas the treated group is nearly symmetric for the middle 50% of its data. Figure 3.32(b) displays the nearly identical distribution of age for the two groups; the only difference is that the treated group has a slightly lower median age and is slightly more variable than is the placebo group. Thus, the two groups appear to have similar age and baseline-seizure distributions prior to the start of the clinical trials.

3.8 RESEARCH STUDY: Controlling for Student Background in the Assessment of Teaching

At the beginning of this chapter, we described a situation faced by many school administrators having a large minority population in their school and/or a large proportion of their students classified as from a low-income family. The implications of such demographics for teacher evaluations based on the performance of their students on standardized reading and math tests generates much controversy in the educational community. The task of achieving goals set by the national *No Child Left Behind* mandate is much more difficult for students from disadvantaged backgrounds. Requiring teachers and administrators from school districts with a high proportion of disadvantaged students to meet the same standards as those for schools with a more advantaged student body is inherently unfair. This type of policy may prove to be counterproductive. It may lead to the alienation of teachers and administrators and the flight of the most qualified and most productive educators from disadvantaged school districts, resulting in a staff with only those educators with an overwhelming commitment to students with a disadvantaged background and/or educators who lack the qualifications to move to the higher-rated schools. A policy that mandates that educators should be held accountable for the success of their students without taking into account the backgrounds of those students is destined for failure.

The data from a medium-sized Florida school district with 22 elementary schools were presented at the beginning of this chapter. The minority status of a student was defined as black or non-black race. In this school district, almost all students are non-Hispanic blacks or whites. Most of the relatively small numbers of Hispanic students are white. Most students of other races are Asian, but they are relatively few in number. They were grouped in the minority category because of the similarity of their test score profiles. Poverty status was based on whether

TABLE 3.21
Summary statistics for reading scores and math scores by grade level

Variable	Grade	N	Mean	St. Dev	Minimum	Q_1	Median	Q_3	Maximum
Math	3	22	171.87	9.16	155.50	164.98	174.65	179.18	186.10
	4	22	189.88	9.64	169.90	181.10	189.45	197.28	206.90
	5	19	206.16	11.14	192.90	197.10	205.20	212.70	228.10
Reading	3	22	171.10	7.46	157.20	164.78	171.85	176.43	183.80
	4	22	185.96	10.20	166.90	178.28	186.95	193.85	204.70
	5	19	205.36	11.04	186.60	199.00	203.30	217.70	223.30
%Minority	3	22	39.43	25.32	12.30	20.00	28.45	69.45	87.40
	4	22	40.22	24.19	11.10	21.25	32.20	64.53	94.40
	5	19	40.42	26.37	10.50	19.80	29.40	64.10	92.60
%Poverty	3	22	58.76	24.60	13.80	33.30	68.95	77.48	91.70
	4	22	54.00	24.20	11.70	33.18	60.55	73.38	91.70
	5	19	56.47	23.48	13.20	37.30	61.00	75.90	92.90

or not the student received a free or reduced lunch subsidy. The math and reading scores are from the Iowa Test of Basic Skills. The number of students by class in each school is given by N in Table 3.21.

The superintendent of schools presented the school board members with the data, and they wanted an assessment of whether poverty and minority status had any effect on the math and reading scores. Just looking at the data presented very little insight in reaching an answer to this question. Using a number of the graphs and summary statistics introduced in this chapter, we will attempt to assist the superintendent in providing insight to the school board concerning the impact of poverty and minority status on student performance.

In order to access the degree of variability in the mean math and reading scores between the 22 schools, a boxplot of the math and reading scores for each of the three grade levels is given in Figure 3.33. There are 22 third- and fourth-grade classes and only 19 fifth-grade classes.

FIGURE 3.33
Boxplot of math and reading scores for each grade

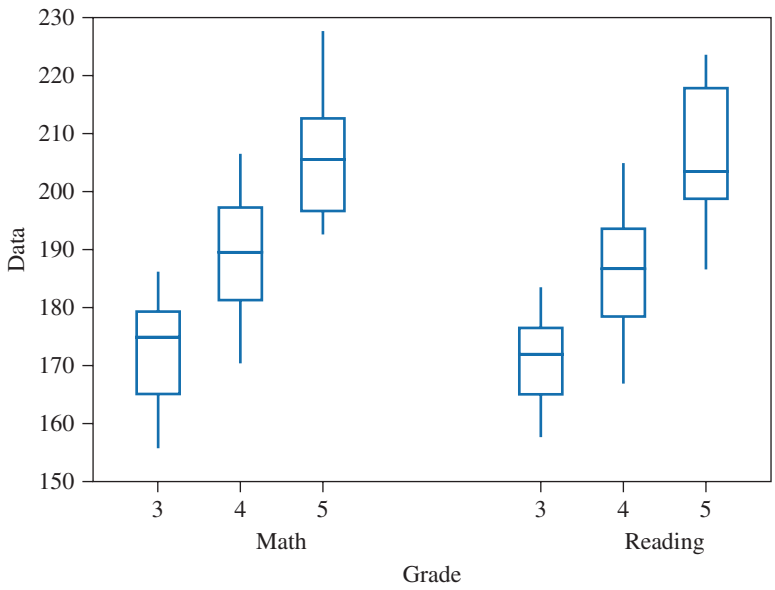
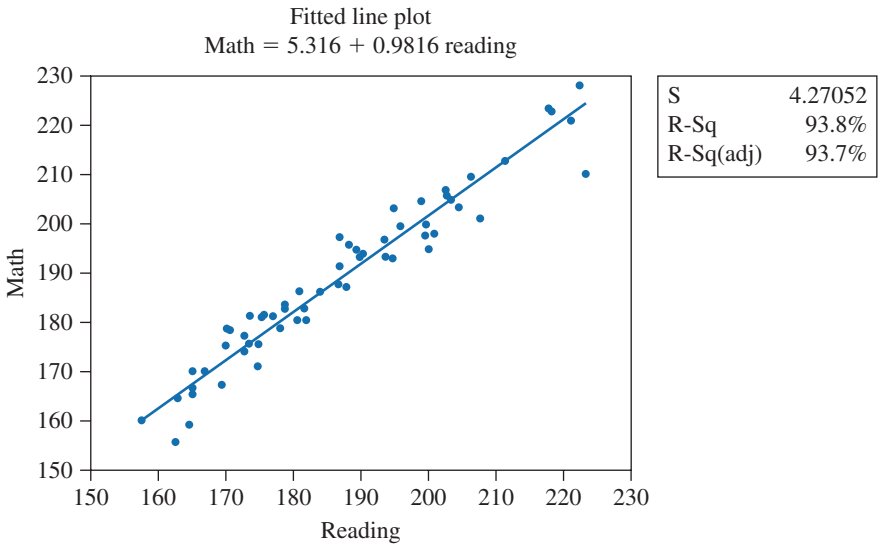


FIGURE 3.34
Scatterplot of reading scores versus math scores



From these plots, we observe that for each of the three grade levels there is a wide variation in mean math and reading scores. However, the level of variability within a grade appears to be about the same for math and reading scores but with a wide level of variability for fourth and fifth graders in comparison to third graders. Furthermore, there is an increase in the median scores from the third to the fifth grades. A detailed summary of the data is given in Table 3.21.

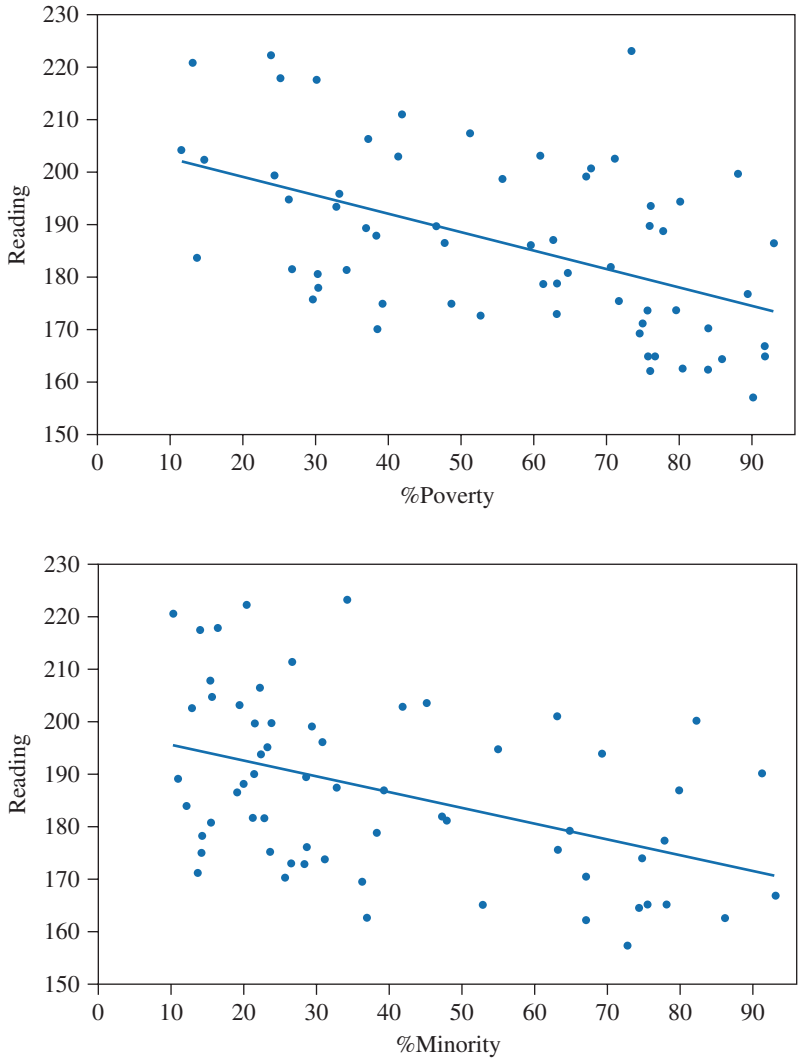
For the third-grade classes, the scores for math and reading had similar ranges: 155 to 185. The range for the 22 schools increased to 170 to 205 for the fourth-grade students in both math and reading. This size of the range for the fifth-grade students was similar to that of the fourth graders: 190 to 225 for both math and reading. Thus, the level of variability in reading and math scores is increasing from third grade to fourth grade to fifth grade. This is confirmed by examining the standard deviations for the three grades. Also, the median scores for both math and reading are increasing across the three grades. The school board then asked the superintendent to identify possible sources of differences in the 22 schools that may help explain the differences in the mean math and reading scores.

In order to simplify the analysis somewhat, it was proposed to analyze just the reading scores because it would appear that the math and reading scores had a similar variation between the 22 schools. To help justify this choice in analysis, a scatterplot of the 63 pairs of math and reading scores (recall there were only 19 fifth-grade classes) was generated (see Figure 3.34). From this plot, we can observe a strong correlation between the reading and math scores for the 63 grades. In fact, the correlation coefficient between math and reading scores is computed to be .97. Thus, there is a very strong relationship between reading and math scores at the 22 schools. The remainder of the analysis will concern the reading scores.

The next step in the process is to examine whether minority or poverty status is associated with the reading scores. Figure 3.35 is a scatterplot of reading versus %poverty and reading versus %minority.

Although there appears to be a general downward trend in reading scores as the levels of %poverty and %minority in the schools increase, there is a wide scattering of individual scores about the fitted line. The correlation between reading

FIGURE 3.35
Scatterplot of reading scores versus %minority and %poverty



and %poverty is −.45, and that between reading and %minority is −.53. However, recall that there is a general upward shift in reading scores from the third grade to the fifth grade. Therefore, a more appropriate plot of the data would be to fit a separate line for each of the three grades. This plot is given in Figure 3.36.

From these plots, we can observe a much stronger association between reading scores and both %poverty and %minority. In fact, if we compute the correlation between the variables separately for each grade level, we will note a dramatic increase in the value of the correlation coefficient. The values are given in Table 3.22.

From Figure 3.36 and the values of the correlation coefficients, we can observe that as the proportion of minority students in the schools increases, there is a steady decline in reading scores. The same pattern is observed with respect to the proportion of students who are classified as being from a low-income family.

What can we conclude from the information presented above? First, it would appear that scores on reading exams tend to decrease as the values of

FIGURE 3.36
Scatterplot of reading scores versus %minority and %poverty with separate lines for each grade

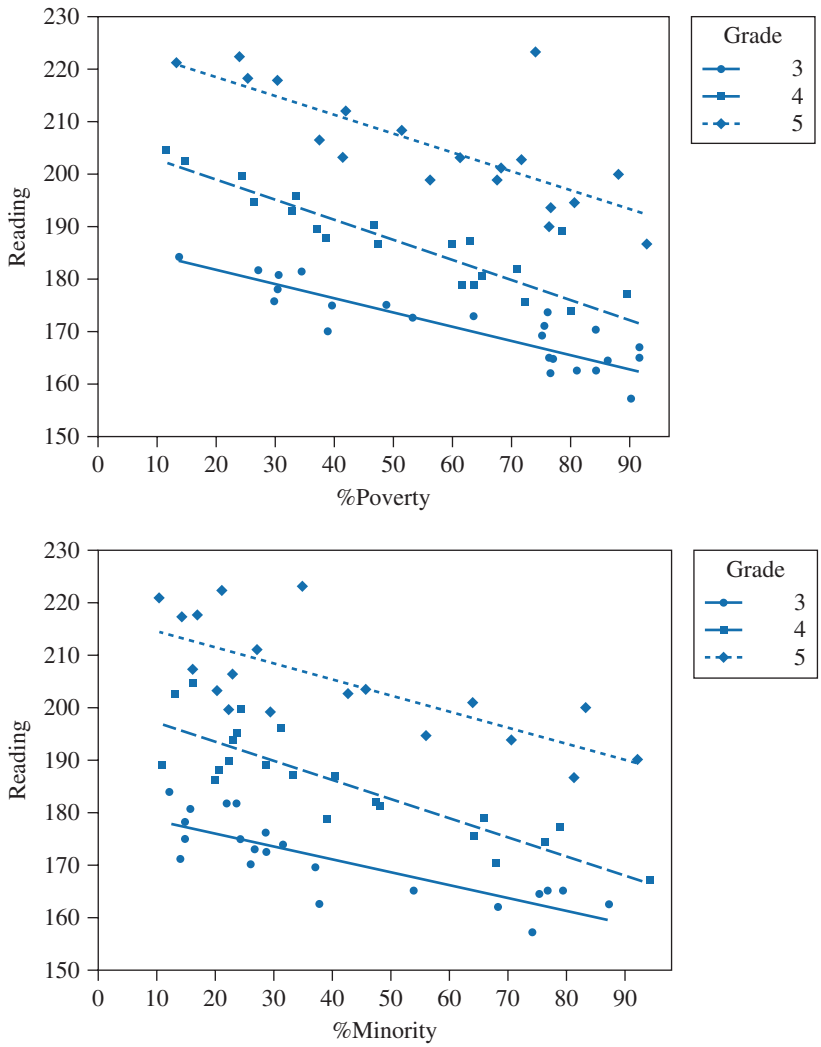


TABLE 3.22
Correlation between reading scores and %poverty and %minority

Correlation Between	3rd Grade	4th Grade	5th Grade
Reading Scores and			
%Minority	−.83	−.87	−.75
%Poverty	−.89	−.92	−.76

%poverty and %minority increase. Thus, we may be inclined to conclude that increasing values of %poverty and %minority *cause* a decline in reading scores and hence that the teachers in schools with high levels of %poverty and %minority should have special considerations when teaching evaluations are conducted. This type of thinking often leads to very misleading conclusions. There may be many other variables involved other than %poverty and %minority that may be impacting the reading scores. To conclude that the high levels %poverty and %minority in a school will often result in low reading scores cannot be supported

by these data. Much more information is needed to reach any conclusion having this type of certainty.

3.9 R Instructions

R Commands for Summarizing Data

Suppose we have two data sets:

Data set 1: 2, 6, 8, 12, −19, 30, 0, −5, 7, 16, 23, 38, −29, 35, 1, −28
Data set 2: 9, 2, −4, 42, 9, 23, −3, −6, 5, 22, −14, 51, 65, 3, −16, −3

The following commands will generate plots of the data and summary statistics:

1. Enter data into R:

   x = c(2, 6, 8, 12, −19, 30, 0, −5, 7, 16, 23, 38, −29, 35, 1, −28)
   y = c(9, 2, −4, 42, 9, 23, −3, −6, 5, 22, −14, 51, 65, 3, −16, −3)

2. Mean: mean(x)
3. Median: median(x)
4. Histogram: hist(x)
5. Stem-and-leaf plot: stem(x)
6. Ordered data: sort(x)
7. Percentiles: quantile(x, seq(0, l, .l))
8. Quantiles at p = .1, .34, .68, .93:

   p = c(.1, .34, .68, .93)
   quantile (x, p)

9. Interquartile range: IQR(x)
10. Variance: var(x)
11. Standard deviation: sd(x)
12. MAD: mad(x)
13. Boxplot: boxplot(x)
14. Scatterplot: plot(x, y)
15. Correlation: cor(x, y)
16. Quantile plot:

    n = length(x)
    i = seq(1 : n)
    u = (i − .5)/n
    s = sort(x)
    plot(u, s)

You can obtain more information about any of the R commands—for example, plot—by just typing **? plot** after the command prompt.

3.10 Summary and Key Formulas

This chapter was concerned with graphical and numerical description of data. The pie chart and bar graph are particularly appropriate for graphically displaying data obtained from a qualitative variable. The frequency and relative frequency

histograms and stem-and-leaf plots are graphical techniques applicable only to quantitative data.

Numerical descriptive measures of data are used to convey a mental image of the distribution of measurements. Measures of central tendency include the mode, the median, and the arithmetic mean. Measures of variability include the range, the interquartile range, the variance, and the standard deviation of a set of measurements.

We extended the concept of data description to summarize the relations between two qualitative variables. Here cross-tabulations were used to develop percentage comparisons. We examined plots for summarizing the relations between quantitative and qualitative variables and between two quantitative variables. Material presented here (namely, summarizing relations among variables) will be discussed and expanded in later chapters on chi-square methods, on the analysis of variance, and on regression.

Key Formulas

Let $y_1, y_2, \ldots, y_n$ be a data set of n values with ordered values $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$

1. Sample median $(\tilde{y})$

   If n is odd, $\tilde{y} = y_{\left(\frac{n+1}{2}\right)}$, middle value

   If n is even, $\tilde{y} = \left[y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}+1\right)}\right]$, average of two middle values

2. Sample median, grouped data

   $$\text{Median} \approx L + \frac{w}{f_m}(.5n - cf_b)$$

3. Sample mean $(\bar{y})$

   $$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

4. Sample mean, grouped data

   $$\bar{y} \approx \frac{\sum_{j=1}^{k} f_j y_j}{n}$$

5. Sample variance

   $$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_i - \bar{y})^2$$

6. Sample variance, grouped data

   $$s^2 \approx \frac{1}{n-1}\sum_{j=1}^{k} f_j(y_j - \bar{y})^2$$

7. Sample standard deviation

   $$s = \sqrt{s^2}$$

8. Sample coefficient of variation

   $$\text{CV} = \frac{s}{|y|}$$

9. Sample MAD

   $$\text{MAD} = \text{median of } (|y_1 - \tilde{y}|, |y_2 - \tilde{y}|, \ldots, |y_n - \tilde{y}|)/.6745$$

10. Sample correlation coefficient

    $$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

3.11 Exercises

3.3 Describing Data on a Single Variable: Graphical Methods

Gov.

3.1 The U.S. government spent more than $3.6 trillion in the 2014 fiscal year. The following table provides broad categories that demonstrate the expenditures of the federal government for domestic and defense programs.

Federal Program	2014 Expenditures (billions of dollars)
National Defense	$612
Social Security	$852
Medicare & Medicaid	$821
National Debt Interest	$253
Major Social-Aid Programs	$562
Other	$532

a. Construct a pie chart for these data.
b. Construct a bar chart for these data.
c. Construct a pie chart and bar chart using percentages in place of dollars.
d. Which of the four charts is more informative to the tax-paying public?

Bus. **3.2** The type of vehicle the U.S public purchases varies depending on many factors. Table 1060 from the ***U.S. Census Bureau, Statistical Abstract of the United States: 2012*** provides the following data. The numbers reported are in thousands of units; that is, 9,300 represents 9,300,000 vehicles sold in 1990.

	Year								
Type of Vehicle	**1990**	**1995**	**2000**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Passenger Car	9,300	8,500	8,852	7,720	7,821	7,618	6,814	5,456	5,729
SUV/Light Truck	4,560	6,340	8,492	9,228	8,683	8,471	6,382	4,945	5,826

a. Construct a graph that would display the changes from 1990 to 2010 in the public's choice in vehicle.
b. Do you observe any trends in the type of vehicle purchased? What factors may be influencing these trends?

Med. **3.3** It has been reported that there has been a change in the type of practice physicians are selecting for their career. In particular, there is concern that there will be a shortage of family practice physicians in future years. The following table contains data on the total number of office-based physicians and the number of those physicians declaring themselves to be family practice physicians. The numbers in the table are given in thousands of physicians. ***(Source: U.S. Census Bureau, Statistical Abstract of the United States: 2002.)***

	Year						
	1980	**1990**	**1995**	**1998**	**1999**	**2000**	**2001**
Family Practice	47.8	57.6	59.9	64.6	66.2	67.5	70.0
Total Office-Based Physicians	271.3	359.9	427.3	468.8	473.2	490.4	514.0

a. Use a bar chart to display the increase in the number of family practice physicians from 1990 to 2001.
b. Calculate the percentage of office-based physicians who are family practice physicians and then display these data in a bar chart.
c. Is there a major difference in the trend displayed by the two bar charts?

Env. **3.4** The regulations of the board of health in a particular state specify that the fluoride level must not exceed 1.5 parts per million (ppm). The 25 measurements given here represent the fluoride levels for a sample of 25 days. Although fluoride levels are measured more than once per day, these data represent the early morning readings for the 25 days sampled.

.75	.86	.84	.85	.97
.94	.89	.84	.83	.89
.88	.78	.77	.76	.82
.72	.92	1.05	.94	.83
.81	.85	.97	.93	.79

a. Determine the range of the measurements.
b. Dividing the range by 7, the number of subintervals selected, and rounding, we have a class interval width of .05. Using .705 as the lower limit of the first interval, construct a frequency histogram.
c. Compute relative frequencies for each class interval and construct a relative frequency histogram. Note that the frequency and relative frequency histograms for these data have the same shape.
d. If one of these 25 days were selected at random, what would be the chance (probability) that the fluoride reading would be greater than .90 ppm? Guess (predict) what proportion of days in the coming year will have a fluoride reading greater than .90 ppm.

Gov. **3.5** The National Highway Traffic Safety Administration has studied the use of rear-seat automobile lap and shoulder seat belts. The number of lives potentially saved with the use of lap and shoulder seat belts is shown for various percentages of use.

Percentage of Use	Lives Saved Wearing	
	Lap Belt Only	Lap and Shoulder Belt
100	529	678
80	423	543
60	318	407
40	212	271
20	106	136
10	85	108

Suggest several different ways to graph these data. Which one seems more appropriate and why?

Soc. **3.6** With the increase in the mobility of the population in the United States and with the increase in home-based employment, there is an inclination to assume that the personal income in the United States will become fairly uniform across the country. The following table provides the per capita personal income for each of the 50 states and the District of Columbia.

Income (thousands of dollars)	Number of States
22.0–24.9	5
25.0–27.9	13
28.0–30.9	16
31.0–33.9	9
34.0–36.9	4
37.0–39.9	2
40.0–42.9	2
Total	51

a. Construct a relative frequency histogram for the income data.
b. Describe the shape of the histogram using the standard terminology of histograms.
c. Would you describe per capita income as being fairly homogenous across the United States?

Med. **3.7** The survival times (in months) for two treatments for patients with severe chronic left-ventricular heart failure are given in the following tables.

Standard Therapy						
4	15	24	10	1	27	31
14	2	16	32	7	13	36
29	6	12	18	14	15	18
6	13	21	20	8	3	24

New Therapy						
5	20	29	15	7	32	36
17	15	19	35	10	16	39
27	14	10	16	12	13	16
9	18	33	30	29	31	27

a. Construct separate relative frequency histograms for the survival times of both the therapies.
b. Compare the two histograms. Does the new therapy appear to generate a longer survival time? Explain your answer.

3.8 Combine the data from the separate therapies in Exercise 3.7 into a single data set, and construct a relative frequency histogram for this combined data set. Does the plot indicate that the data are from two separate populations? Explain your answer.

Gov. **3.9** Liberal members of Congress have asserted that the U.S. government has been expending an increasing portion of the nation's resources on the military and intelligence agencies since 1960. The following table contains the outlays (in billion of dollars) for the Defense Department and associated intelligence agencies since 1960. The data are also given as a percentage of gross national product (%GNP).

Year	Expenditure	%GNP	Year	Expenditure	%GNP
1960	48	9.3	1996	266	3.5
1970	81	8.1	1997	271	3.3
1980	134	4.9	1998	269	3.1
1981	158	5.2	1999	275	3.0
1982	185	5.8	2000	295	3.0
1983	210	6.1	2001	306	3.0
1984	227	6.0	2002	349	3.3
1985	253	6.1	2003	376	3.3
1986	273	6.2	2004	456	3.8
1987	282	6.1	2005	495	3.9
1988	290	5.9	2006	522	3.9
1989	304	5.7	2007	551	3.9
1990	299	5.2	2008	616	4.3
1991	273	4.6	2009	661	4.7
1992	298	4.8	2010	694	4.7
1993	291	4.4	2011	768	5.1
1994	282	4.1	2012	738	4.7
1995	272	3.7			

Source: U.S. Census Bureau, Statistical Abstract of the United States, 2012.

a. Plot the defense expenditures time series data and describe any trends across the period from 1960 to 2012.
b. Plot the % GNP time series data and describe any trends across the period from 1960 to 2012.
c. Do the two time series have similar trends? Do either of the plots support the assertions by the liberal members of Congress?
d. What factors, domestic or international, do you think may have had an influence on your observed trends?

Soc. **3.10** The following table presents homeownership rates, in percentages, by state for the years 1985, 1996, and 2002. These values represent the proportion of homes owned by the occupant to the total number of occupied homes.

State	1985	1996	2002	State	1985	1996	2002
Alabama	70.4	71.0	73.5	Montana	66.5	68.6	69.3
Alaska	61.2	62.9	67.3	Nebraska	68.5	66.8	68.4
Arizona	64.7	62.0	65.9	Nevada	57.0	61.1	65.5
Arkansas	66.6	66.6	70.2	New Hampshire	65.5	65.0	69.5
California	54.2	55.0	58.0	New Jersey	62.3	64.6	67.2
Colorado	63.6	64.5	69.1	New Mexico	68.2	67.1	70.3
Connecticut	69.0	69.0	71.6	New York	50.3	52.7	55.0
Delaware	70.3	71.5	75.6	North Carolina	68.0	70.4	70.0
Dist. of Columbia	37.4	40.4	44.1	North Dakota	69.9	68.2	69.5
Florida	67.2	67.1	68.7	Ohio	67.9	69.2	72.0
Georgia	62.7	69.3	71.7	Oklahoma	70.5	68.4	69.4
Hawaii	51.0	50.6	57.4	Oregon	61.5	63.1	66.2
Idaho	71.0	71.4	73.0	Pennsylvania	71.6	71.7	74.0
Illinois	60.6	68.2	70.2	Rhode Island	61.4	56.6	59.6
Indiana	67.6	74.2	75.0	South Carolina	72.0	72.9	77.3
Iowa	69.9	72.8	73.9	South Dakota	67.6	67.8	71.5
Kansas	68.3	67.5	70.2	Tennessee	67.6	68.8	70.1
Kentucky	68.5	73.2	73.5	Texas	60.5	61.8	63.8
Louisiana	70.2	64.9	67.1	Utah	71.5	72.7	72.7
Maine	73.7	76.5	73.9	Vermont	69.5	70.3	70.2
Maryland	65.6	66.9	72.0	Virginia	68.5	68.5	74.3
Massachusetts	60.5	61.7	62.7	Washington	66.8	63.1	67.0
Michigan	70.7	73.3	76.0	West Virginia	75.9	74.3	77.0
Minnesota	70.0	75.4	77.3	Wisconsin	63.8	68.2	72.0
Mississippi	69.6	73.0	74.8	Wyoming	73.2	68.0	72.8
Missouri	69.2	70.2	74.6				

Source: U.S. Bureau of the Census, http://www.census.gov/ftp/pub/hhes/www/hvs.html.

a. Construct relative frequency histogram plots for the homeownership data given in the table for the years 1985, 1996, and 2002.
b. What major differences exist among the plots for the three years?
c. Why do you think the plots have changed over these 17 years?
d. How could Congress use the information in these plots for writing tax laws that allow major tax deductions for homeownership?

3.11 Construct a stem-and-leaf plot for the data of Exercise 3.10.

3.12 Describe the shape of the stem-and-leaf plot and histogram for the homeownership data in Exercises 3.10 and 3.11, using the terms *modality, skewness,* and *symmetry* in your description.

Bus. **3.13** A supplier of high-quality audio equipment for automobiles accumulates monthly sales data on speakers and receiver–amplifier units for 5 years. The data (in thousands of units per

month) are shown in the following table. Plot the sales data. Do you see any overall trend in the data? Do there seem to be any cyclic or seasonal effects?

Year	J	F	M	A	M	J	J	A	S	O	N	D
1	101.9	93.0	93.5	93.9	104.9	94.6	105.9	116.7	128.4	118.2	107.3	108.6
2	109.0	98.4	99.1	110.7	100.2	112.1	123.8	135.8	124.8	114.1	114.9	112.9
3	115.5	104.5	105.1	105.4	117.5	106.4	118.6	130.9	143.7	132.2	120.8	121.3
4	122.0	110.4	110.8	111.2	124.4	112.4	124.9	138.0	151.5	139.5	127.7	128.0
5	128.1	115.8	116.0	117.2	130.7	117.5	131.8	145.5	159.3	146.5	134.0	134.2

3.4 Describing Data on a Single Variable: Measures of Central Tendency

Basic **3.14** Compute the mean, median, and mode for the following data:

155	25	30	52	142	35	51	26	2	23
270	74	29	29	29	29	51	83	9	69

Basic **3.15** Compute the mean, median, and mode for the following data:

35	81	96	45	109	126	71	15	8	79	56
73	58	17	82	29	58	68	24	5	24	

Basic **3.16** Refer to the data in Exercise 3.15 with the measurements 109 and 126 replaced by 378 and 517. Recompute the mean, median, and mode. Discuss the impact of these extreme measurements on the three measures of central tendency.

Basic **3.17** Compute a 10% trimmed mean for the data sets in Exercises 3.15 and 3.16. Do the extreme values in Exercise 3.16 affect the 10% trimmed mean? Would a 5% trimmed mean be as affected by the two extreme values as the 10% trimmed mean?

Basic **3.18** A data set of 75 values is summarized in the following frequency table. Estimate the mean, median, and mode for the 75 data values using the summarized data.

Class Interval	Frequency
2.0–4.9	9
5.0–7.9	19
8.0–10.9	27
11.0–13.9	10
14.0–16.9	5
17.0–19.9	3
20.0–22.9	2

Engin. **3.19** A study of the reliability of buses ["Large Sample Simultaneous Confidence Intervals for the Multinominal Probabilities on Transformations of the Cell Frequencies," ***Technometrics (1980) 22:588***] examined the reliability of 191 buses. The distance traveled (in 1,000s of miles) prior to the first major motor failure was classified into intervals. A modified form of the table follows.

Distance Traveled (1,000s of miles)	Frequency
0–20.0	6
20.1–40.0	11
40.1–60.0	16
60.1–100.0	59
100.1–120.0	46
120.1–140.0	33
140.1–160.0	16
160.1–200.0	4

a. Sketch the relative frequency histogram for the distance data and describe its shape.
b. Estimate the mode, median, and mean for the distance traveled by the 191 buses.
c. What does the relationship among the three measures of center indicate about the shape of the histogram for these data?
d. Which of the three measures would you recommend as the most appropriate representative of the distance traveled by one of the 191 buses? Explain your answer.

Med. **3.20** In a study of 1,329 American men reported in **American Statistician *[(1974) 28:115–122],*** the men were classified by serum cholesterol and blood pressure. The group of 408 men who had blood pressure readings less than 127 mm Hg were then classified according to their serum cholesterol level.

Serum Cholesterol (mg/100cc)	Frequency
0.0–199.9	119
200.0–219.9	88
220.0–259.9	127
greater than 259	74

a. Estimate the mode, median, and mean for the serum cholesterol readings (if possible).
b. Which of the three summary statistics is most informative concerning a typical serum cholesterol level for the group of men? Explain your answer.

Env. **3.21** The ratio of DDE (related to DDT) to PCB concentrations in bird eggs has been shown to have a number of biological implications. The ratio is used as an indication of the movement of contamination through the food chain. The paper ***"The Ratio of DDE to PCB Concentrations in Great Lakes Herring Gull Eggs and Its Use in Interpreting Contaminants Data" [*Journal of Great Lakes Research *(1998) 24(1):12–31]*** reports the following ratios for eggs collected at 13 study sites from the five Great Lakes. The eggs were collected from both terrestrial- and aquatic-feeding birds.

	DDE to PCB Ratio										
Terrestrial Feeders	76.50	6.03	3.51	9.96	4.24	7.74	9.54	41.70	1.84	2.50	1.54
Aquatic Feeders	0.27	0.61	0.54	0.14	0.63	0.23	0.56	0.48	0.16	0.18	

a. Compute the mean and median for the 21 ratios, ignoring the type of feeder.
b. Compute the mean and median separately for each type of feeder.
c. Using your results from parts (a) and (b), comment on the relative sensitivity of the mean and median to extreme values in a data set.

d. Which measure, mean or median, would you recommend as the more appropriate measure of the DDE to PCB level for both types of feeders? Explain your answer.

Med. **3.22** A study of the survival times, in days, of skin grafts on burn patients was examined by Woolson and Lachenbruch *[Biometrika (1980) 67:597–606]*. Two of the patients left the study prior to the failure of their grafts. The survival time for these individuals is some number greater than the reported value.

Survival time (days): 37, 19, 57*, 93, 16, 22, 20, 18, 63, 29, 60*

(The "*" indicates that the patient left the study prior to failure of the graft; values given are for the day the patient left the study.)

a. Calculate the measures of center (if possible) for the 11 patients.

b. If the survival times of the two patients who left the study were obtained, how would these new values change the values of the summary statistics calculated in (a)?

Engin. **3.23** A study of the reliability of diesel engines was conducted on 14 engines. The engines were run in a test laboratory. The time (in days) until the engine failed is given here. The study was terminated after 300 days. For those engines that did not fail during the study period, an asterisk is placed by the number 300. Thus, for these engines, the time to failure is some value greater than 300.

Failure time (days): 130, 67, 300*, 234, 90, 256, 87, 120, 201, 178, 300*, 106, 289, 74

a. Calculate the measures of center for the 14 engines.

b. What are the implications of computing the measures of center when some of the exact failure times are not known?

Gov. **3.24** Effective tax rates (per $100) on residential property for three groups of large cities, ranked by residential property tax rate, are shown in the following table.

Group 1	Rate	Group 2	Rate	Group 3	Rate
Detroit, MI	4.10	Burlington, VT	1.76	Little Rock, AR	1.02
Milwaukee, WI	3.69	Manchester, NH	1.71	Albuquerque, NM	1.01
Newark, NJ	3.20	Fargo, ND	1.62	Denver, CO	.94
Portland, OR	3.10	Portland ME	1.57	Las Vegas, NV	.88
Des Moines, IA	2.97	Indianapolis, IN	1.57	Oklahoma City, OK	.81
Baltimore, MD	2.64	Wilmington, DE	1.56	Casper, WY	.70
Sioux Falls, IA	2.47	Bridgeport, CT	1.55	Birmingham, AL	.70
Providence, RI	2.39	Chicago, IL	1.55	Phoenix, AZ	.68
Philadelphia, PA	2.38	Houston, TX	1.53	Los Angeles, CA	.64
Omaha, NE	2.29	Atlanta, GA	1.50	Honolulu, HI	.59

Source: Government of the District of Columbia, Department of Finance and Revenue,* Tax Rates and Tax Burdens in the District of Columbia: A Nationwide Comparison *(annual).

a. Compute the mean, median, and mode separately for the three groups.

b. Compute the mean, median, and mode for the complete set of 30 measurements.

c. What measure or measures best summarize the center of these distributions? Explain.

3.25 Refer to Exercise 3.24. Average the three group means, the three group medians, and the three group modes, and compare your results to those of part (b). Comment on your findings.

3.5 Describing Data on a Single Variable: Measures of Variability

Engin. **3.26** Pushing economy and wheelchair-propulsion technique were examined for eight wheelchair racers on a motorized treadmill in a paper by Goosey and Campbell *[Adapted Physical Activity Quarterly (1998) 15:36–50]*. The eight racers had the following years of racing experience:

Racing experience (years): 6, 3, 10, 4, 4, 2, 4, 7

a. Verify that the mean years of experience is 5 years. Does this value appear to adequately represent the center of the data set?
b. Verify that $\sum_i (y - \bar{y})^2 = \sum_i (y - 5)^2 = 46$.
c. Calculate the sample variance and standard deviation for the experience data. How would you interpret the value of the standard deviation relative to the sample mean?

3.27 In the study described in Exercise 3.26, the researchers also recorded the ages of the eight racers.

Age (years): 39, 38, 31, 26, 18, 36, 20, 31

a. Calculate the sample standard deviation of the eight racers' ages.
b. Why would you expect the standard deviation of the racers' ages to be larger than the standard deviation of their years of experience?

Engin. **3.28** For the data in Exercises 3.26 and 3.27,
a. Calculate the coefficient of variation (CV) for both the racers' ages and their years of experience. Are the two CVs relatively the same? Compare their relative sizes to the relative sizes of their standard deviations.
b. Estimate the standard deviations for both the racers' ages and their years of experience by dividing the ranges by 4. How close are these estimates to the standard deviations calculated in Exercises 3.26 and 3.27?

Med. **3.29** The treatment times (in minutes) for patients at a health clinic are as follows:

21	20	31	24	15	21	24	18	33	8
26	17	27	29	24	14	29	41	15	11
13	28	22	16	12	15	11	16	18	17
29	16	24	21	19	7	16	12	45	24
21	12	10	13	20	35	32	22	12	10

Construct the quantile plot for the treatment times for the patients at the health clinic.
a. Find the 25th percentile for the treatment times and interpret this value.
b. The health clinic advertises that 90% of all its patients have a treatment time of 40 minutes or less. Do the data support this claim?

Env. **3.30** To assist in estimating the amount of lumber in a tract of timber, an owner decided to count the number of trees with diameters exceeding 12 inches in randomly selected 50×50-foot squares. Seventy 50×50 squares were randomly selected from the tract and the number of trees (with diameters in excess of 12 inches) was counted for each. The data are as follows:

7	8	6	4	9	11	9	9	9	10
9	8	11	5	8	5	8	8	7	8
3	5	8	7	10	7	8	9	8	11
10	8	9	8	9	9	7	8	13	8
9	6	7	9	9	7	9	5	6	5
6	9	8	8	4	4	7	7	8	9
10	2	7	10	8	10	6	7	7	8

a. Construct a relative frequency histogram to describe these data.
b. Calculate the sample mean $\bar{y}$ as an estimate of μ, the mean number of timber trees with diameter exceeding 12 inches for all 50×50 squares in the tract.
c. Calculate s for the data. Construct the intervals $(\bar{y} \pm s)$, $(\bar{y} \pm 2s)$, and $(\bar{y} \pm 3s)$. Count the percentages of squares falling in each of the three intervals, and compare these percentages with the corresponding percentages given by the Empirical Rule.

Bus. **3.31** ***Consumer Reports*** in its June 1998 issue reports on the typical daily room rate at six luxury and nine budget hotels. The room rates are given in the following table.

Luxury Hotel	$175	$180	$120	$150	$120	$125			
Budget Hotel	$50	$50	$49	$45	$36	$45	$50	$50	$40

a. Compute the mean and standard deviation of the room rates for both luxury and budget hotels.
b. Verify that luxury hotels have a more variable room rate than budget hotels.
c. Give a practical reason why the luxury hotels are more variable than the budget hotels.
d. Might another measure of variability be better to compare luxury and budget hotel rates? Explain.

Env. **3.32** Many marine phanerogam species are highly sensitive to changes in environmental conditions. In the article ***"Posidonia oceanica: A Biological Indicator of Past and Present Mercury Contamination in the Mediterranean Sea" [Marine Environmental Research, March 1998 45:101–111]***, the researchers report the mercury concentrations over a period of about 20 years at several locations in the Mediterranean Sea. Samples of *Posidonia oceanica* were collected by scuba diving at a depth of 10 meters. For each site, 45 orthotropic shoots were sampled and the mercury concentration was determined. The average mercury concentration is recorded in the following table for each of the sampled years.

Mercury Concentration (ng/g dry weight)		
Year	**Site 1 Calvi**	**Site 2 Marseilles-Coriou**
1992	14.8	70.2
1991	12.9	160.5
1990	18.0	102.8
1989	8.7	100.3
1988	18.3	103.1
1987	10.3	129.0
1986	19.3	156.2
1985	12.7	117.6
1984	15.2	170.6
1983	24.6	139.6
1982	21.5	147.8
1981	18.2	197.7
1980	25.8	262.1
1979	11.0	123.3
1978	16.5	363.9
1977	28.1	329.4
1976	50.5	542.6
1975	60.1	369.9
1974	96.7	705.1
1973	100.4	462.0
1972	*	556.1
1971	*	461.4
1970	*	628.8
1969	*	489.2

a. Generate a time-series plot of the mercury concentrations and place lines for both sites on the same graph. Comment on any trends in the lines across the years of data. Are the trends similar for both sites?
b. Select the most appropriate measure of center for the mercury concentrations. Compare the centers for the two sites.
c. Compare the variabilities of the mercury concentrations at the two sites. Use the CV in your comparison, and explain why it is more appropriate than using the standard deviations.
d. When comparing the centers and variabilities of the two sites, should the years 1969–1972 be used for site 2?

3.6 The Boxplot

Med. **3.33** Construct a boxplot for the following measurements:

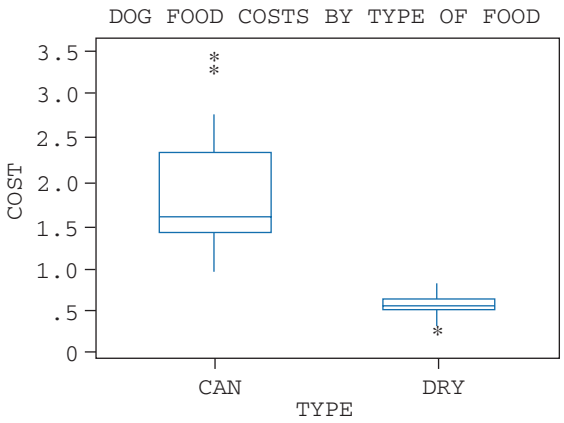
33, 31, 19, 25, 23, 27, 11, 9, 29, 3, 17, 9, 2, 5, 8, 2, 9, 1, 3

Med. **3.34** The following data are the resting pulse rates for 30 randomly selected individuals who were participants at a 10K race.

49	40	59	56	55	70	49	59	55	49	58	54	55	72	51
54	56	55	65	57	61	41	52	60	49	57	46	55	63	55

a. Construct a stem-and-leaf plot of the pulse rates.
b. Construct a boxplot of the pulse rates.
c. Describe the shape of the distribution of the pulse rates.
d. The boxplot provides information about the distribution of pulse rates for what population?

Bus. **3.35** ***Consumer Reports*** in its May 1998 issue provides cost per daily feeding for 28 brands of dry dog food and 23 brands of canned dog food. Using the Minitab computer program, the following side-by-side boxplot for these data was created.



a. From these graphs, determine the median, lower quartile, and upper quartile for the daily costs of both dry and canned dog food.
b. Comment on the similarities and differences in the distributions of daily costs for the two types of dog food.

3.7 Summarizing Data from More Than One Variable: Graphs and Correlation

Soc. **3.36** For the homeownership rates given in Exercise 3.10, construct separate boxplots for the years 1985, 1996, and 2002.

a. Describe the distributions of homeownership rates for each of the 3 years.
b. Compare the descriptions given in part (a) to the descriptions given in Exercise 3.10.

Soc. **3.37** Compute the mean, median, and standard deviation for the homeownership rates given in Exercise 3.10.

a. Compare the mean and median for the 3 years of data. Which value, mean or median, is more appropriate for these data sets? Explain your answers.
b. Compare the degrees of variability in homeownership rates over the 3 years.

Soc. **3.38** For the boxplots constructed for the homeownership rates given in Exercise 3.36, place the three boxplots on the same set of axes.

a. Use this side-by-side boxplot to discuss changes in the median homeownership rate over the 3 years.
b. Use this side-by-side boxplot to discuss changes in the variation in these rates over the 3 years.
c. Are there any states that have extremely low homeownership rates?
d. Are there any states that have extremely high homeownership rates?

Soc. **3.39** In the paper ***"Demographic Implications of Socioeconomic Transition Among the Tribal Populations of Manipur, India"*** **[Human Biology *(1998) 70(3):597–619]***, the authors describe the tremendous changes that have taken place in all the tribal populations of Manipur, India, since the beginning of the twentieth century. The tribal populations of Manipur are in the process of socioeconomic transition from a traditional subsistence economy to a market-oriented economy. The following table displays the relation between literacy level and subsistence group for a sample of 614 married men and women in Manipur, India.

	Literacy Level		
Subsistence Group	**Illiterate**	**Primary Schooling**	**At Least Middle School**
Shifting Cultivators	114	10	45
Settled Agriculturists	76	2	53
Town Dwellers	93	13	208

a. Graphically depict the data in the table using a stacked bar graph.
b. Do a percentage comparison based on the row and column totals. What conclusions do you reach with respect to the relation between literacy and subsistence group?

Engin. **3.40** In the manufacture of soft contact lenses, the power (the strength) of the lens needs to be very close to the target value. In the paper ***"An ANOM-Type Test for Variances from Normal Populations"*** **[Technometrics *(1997) 39:274–283]***, a comparison of several suppliers is made relative to the consistency of the power of the lens. The following table contains the deviations from the target power value of lenses produced using materials from three different suppliers:

Supplier	**Deviations from Target Power Value**								
1	189.9	191.9	190.9	183.8	185.5	190.9	192.8	188.4	189.0
2	156.6	158.4	157.7	154.1	152.3	161.5	158.1	150.9	156.9
3	218.6	208.4	187.1	199.5	202.0	211.1	197.6	204.4	206.8

a. Compute the mean and standard deviation for the deviations of each supplier.
b. Plot the sample deviation data.
c. Describe the deviation from specified power for the three suppliers.
d. Which supplier appears to provide material that produces lenses having power closest to the target value?

Bus. **3.41** The federal government keeps a close watch on money growth versus targets that have been set for that growth. We list two measures of the money supply in the United States, M2 (private checking deposits, cash, and some savings) and M3 (M2 plus some investments), which are given here for 20 consecutive months.

	Money Supply (in trillions of dollars)			Money Supply (in trillions of dollars)	
Month	M2	M3	Month	M2	M3
1	2.25	2.81	11	2.43	3.05
2	2.27	2.84	12	2.42	3.05
3	2.28	2.86	13	2.44	3.08
4	2.29	2.88	14	2.47	3.10
5	2.31	2.90	15	2.49	3.10
6	2.32	2.92	16	2.51	3.13
7	2.35	2.96	17	2.53	3.17
8	2.37	2.99	18	2.53	3.18
9	2.40	3.02	19	2.54	3.19
10	2.42	3.04	20	2.55	3.20

a. Would a scatterplot describe the relation between M2 and M3?
b. Construct a scatterplot. Is there an obvious relation?

3.42 Refer to Exercise 3.41. What other data plot might be used to describe and summarize these data? Make the plot and interpret your results.

Supplementary Exercises

Env. **3.43** To control the risk of severe core damage during a commercial nuclear power station blackout accident, the reliability of the emergency diesel generators in starting on demand must be maintained at a high level. The paper ***"Empirical Bayes Estimation of the Reliability of Nuclear-Power Emergency Diesel Generators"*** *[Technometrics (1996) 38:11–23]* contains data on the failure history of seven nuclear power plants. The following data are the number of successful demands between failures for the diesel generators at one of these plants from 1982 to 1988.

28	50	193	55	4	7	147	76	10	0	10	84	0	9	1	0	62
26	15	226	54	46	128	4	105	40	4	273	164	7	55	41	26	6

(*Note:* The failure of the diesel generator does not necessarily result in damage to the nuclear core because all nuclear power plants have several emergency diesel generators.)

a. Calculate the mean and median of the successful demands between failures.
b. Which measure appears to best represent the center of the data?
c. Calculate the range and standard deviation, s.
d. Use the range approximation to estimate s. How close is the approximation to the true value?
e. Construct the intervals

$$\bar{y} \pm s \qquad \bar{y} \pm 2s \qquad \bar{y} \pm 3s$$

Count the number of demands between failures falling in each of the three intervals. Convert these numbers to percentages and compare your results to the Empirical Rule.
f. Why do you think the Empirical Rule and your percentages do not match well?

Edu. **3.44** The College of Dentistry at the University of Florida has made a commitment to develop its entire curriculum around the use of self-paced instructional materials such as videotapes, slide

tapes, and syllabi. It is hoped that each student will proceed at a pace commensurate with his or her ability and that the instructional staff will have more free time for personal consultation in student–faculty interaction. One such instructional module was developed and tested on the first 50 students proceeding through the curriculum. The following measurements represent the number of hours it took these students to complete the required modular material.

16	8	33	21	34	17	12	14	27	6
33	25	16	7	15	18	25	29	19	27
5	12	29	22	14	25	21	17	9	4
12	15	13	11	6	9	26	5	16	5
9	11	5	4	5	23	21	10	17	15

a. Calculate the mode, the median, and the mean for these recorded completion times.
b. Guess the value of s.
c. Compute s by using the shortcut formula and compare your answer to that of part (b).
d. Would you expect the Empirical Rule to describe adequately the variability of these data? Explain.

Bus. **3.45** The February 1998 issue of ***Consumer Reports*** provides data on the price of 24 brands of paper towels. The prices are given in both cost per roll and cost per sheet because the brands had varying numbers of sheets per roll.

Brand	Price per Roll	Number of Sheets per Roll	Cost per Sheet
1	1.59	50	.0318
2	0.89	55	.0162
3	0.97	64	.0152
4	1.49	96	.0155
5	1.56	90	.0173
6	0.84	60	.0140
7	0.79	52	.0152
8	0.75	72	.0104
9	0.72	80	.0090
10	0.53	52	.0102
11	0.59	85	.0069
12	0.89	80	.0111
13	0.67	85	.0079
14	0.66	80	.0083
15	0.59	80	.0074
16	0.76	80	.0095
17	0.85	85	.0100
18	0.59	85	.0069
19	0.57	78	.0073
20	1.78	180	.0099
21	1.98	180	.0100
22	0.67	100	.0067
23	0.79	100	.0079
24	0.55	90	.0061

a. Compute the standard deviation for both the price per roll and the price per sheet.
b. Which is more variable, price per roll or price per sheet?
c. In your comparison in part (b), should you use s or CV? Justify your answer.

3.46 Refer to Exercise 3.45. Use a scatterplot to plot the price per roll and number of sheets per roll.

- **a.** Do the 24 points appear to fall on a straight line?
- **b.** If not, is there any other relation between the two prices?
- **c.** What factors may explain why the ratio of price per roll to number of sheets is not a constant?

3.47 Refer to Exercise 3.45. Construct boxplots for both price per roll and number of sheets per roll. Are there any "unusual" brands in the data?

Env. **3.48** The paper ***"Conditional Simulation of Waste-Site Performance" [Technometrics (1994) 36: 129–161]*** discusses the evaluation of a pilot facility for demonstrating the safe management, storage, and disposal of defense-generated, radioactive, transuranic waste. Researchers have determined that one potential pathway for release of radionuclides is through contaminant transport in groundwater. Recent focus has been on the analysis of transmissivity, a function of the properties and the thickness of an aquifer that reflects the rate at which water is transmitted through the aquifer. The following table contains 41 measurements of transmissivity, T, made at the pilot facility.

9.354	6.302	24.609	10.093	0.939	354.81	15399.27	88.17	1253.43	0.75	312.10
1.94	3.28	1.32	7.68	2.31	16.69	2772.68	0.92	10.75	0.000753	
1.08	741.99	3.23	6.45	2.69	3.98	2876.07	12201.13	4273.66	207.06	
2.50	2.80	5.05	3.01	462.38	5515.69	118.28	10752.27	956.97	20.43	

- **a.** Draw a relative frequency histogram for the 41 values of T.
- **b.** Describe the shape of the histogram.
- **c.** When the relative frequency histogram is highly skewed to the right, the Empirical Rule may not yield very accurate results. Verify this statement for the data given.
- **d.** Data analysts often find it easier to work with mound-shaped relative frequency histograms. A transformation of the data will sometimes achieve this shape. Replace the given 41 T values with the logarithm base 10 of the values and reconstruct the relative frequency histogram. Is the shape more mound-shaped than the original data? Apply the Empirical Rule to the transformed data, and verify that it yields more accurate results than it did with the original data.

Soc. **3.49** A random sample of 90 standard metropolitan statistical areas (SMSAs) was studied to obtain information on murder rates. The murder rates (number of murders per 100,000 people) were recorded, and these data are summarized in the following frequency table.

Class Interval	f_i	Class Interval	f_i
−.5–1.5	2	13.5–15.5	9
1.5–3.5	18	15.5–17.5	4
3.5–5.5	15	17.5–19.5	2
5.5–7.5	13	19.5–21.5	1
7.5–9.5	9	21.5–23.5	1
9.5–11.5	8	23.5–25.5	1
11.5–13.5	7		

Construct a relative frequency histogram for these data.

3.50 Refer to the data of Exercise 3.49.

- **a.** Estimate the sample mean, sample median, and sample mode.
- **b.** Which measure of center would you recommend using as a measure of the center of the distribution for the murder rates?

3.51 Refer to the data of Exercise 3.49.

a. Estimate the interquartile range and the sample standard deviation.

b. Which measure of variation would you recommend using as a measure of the variation in the murder rates?

c. Identify the population to which the measures of center and variation would be reasonable estimators.

Gov. **3.52** Refer to the homeownership data in Exercise 3.10.

a. Construct a quantile plot for each of the 3 years of data. Place these plots on the same set of axes.

b. Congress wants to develop special programs for those states having low homeownership percentages. Which states fell into the lower 10th percentile of homeownership during 2002?

c. Was there a change in the states falling into the 10th percentile during the 3 years, 1985, 1996, and 2002?

Gov. **3.53** Refer to the homeownership data in Exercise 3.10.

a. Compute mean and median homeownership percentages during the 3 years.

b. Which measure best represents the average homeownership percentage during each of the 3 years?

c. Compute standard deviation and MAD homeownership percentage during the 3 years.

d. Which measure best represents the variation in homeownership percentages across the U.S during each of the 3 years?

e. Describe the change in the percentage of homes owned by the occupant over the 3 years.

Engin. **3.54** The ***Insurance Institute for Highway Safety*** published data on the total damage suffered by compact automobiles in a series of controlled, low-speed collisions. The data, in dollars, with brand names removed are as follows:

361	393	430	543	566	610	763	851
886	887	976	1,039	1,124	1,267	1,328	1,415
1,425	1,444	1,476	1,542	1,544	2,048	2,197	

a. Draw a histogram of the data using six or seven categories.

b. On the basis of the histogram, what would you guess the mean to be?

c. Calculate the median and mean.

d. What does the relation between the mean and median indicate about the shape of the data?

Soc. **3.55** Data are collected on the weekly expenditures of a sample of urban households on food (including restaurant expenditures). The data, obtained from diaries kept by each household, are grouped by number of members of the household. The expenditures are as follows:

1 member:	67	62	168	128	131	118	80	53	99	68	
	76	55	84	77	70	140	84	65	67	183	
2 members:	129	116	122	70	141	102	120	75	114	81	106	95
	94	98	85	81	67	69	119	105	94	94	92	
3 members:	79	99	171	145	86	100	116	125				
	82	142	82	94	85	191	100	116				
4 members:	139	251	93	155	158	114	108					
	111	106	99	132	62	129	91					
5+ members:	121	128	129	140	206	111	104	109	135	136		

a. Compute the mean expenditure separately for each of the five groups.

b. Combine the five data sets into a single data set and then compute the mean expenditure.

c. Describe a method by which the mean for the combined data set could be obtained from the five individual means.

e. Describe the relation (if any) among the mean expenditures for the five groups.

3.56 Refer to the data of Exercise 3.55.

a. Compute the standard deviation in the expenditures separately for each of the five groups.

b. Combine the five data sets into a single data set and then compute the standard deviation in expenditures.

c. Describe a method by which the standard deviation for the combined data set could be obtained from the five individual standard deviations.

d. Which group appears to have the largest variability in expenditures?

Gov. **3.57** Federal authorities have destroyed considerable amounts of wild and cultivated marijuana plants. The following table shows the number of plants destroyed and the number of arrests for a 12-month period for 15 states.

State	Plants	Arrests
1	110,010	280
2	256,000	460
3	665	6
4	367,000	66
5	4,700,000	15
6	4,500	8
7	247,000	36
8	300,200	300
9	3,100	9
10	1,250	4
11	3,900,200	14
12	68,100	185
13	450	5
14	2,600	4
15	205,844	33

a. Discuss the appropriateness of using the sample mean to describe these two variables.

b. Compute the sample mean, 10% trimmed mean, and 20% trimmed mean. Which trimmed mean seems more appropriate for each variable? Why?

c. Does there appear to be a relation between the number of plants destroyed and the number of arrests? How might you examine this question? What other variable(s) might be related to the number of plants destroyed?

Bus. **3.58** The most widely reported index of the performance of the New York Stock Exchange (NYSE) is the Dow Jones Industrial Average (DJIA). This index is computed from the stock prices of 30 companies. When the DJIA was invented in 1896, the index was the average price of 12 stocks. The index was modified over the years as new companies were added and dropped from the index and was also altered to reflect when a company splits its stock. The closing ***New York Stock Exchange (NYSE)*** prices for the 30 components (as of June 19, 2014) of the DJIA are given in the following table.

a. Compute the average price of the 30 stock prices in the DJIA.

b. The DJIA is no longer an average; the name includes the word "average" only for historical reasons. The index is computed by summing the stock prices and dividing by a constant, which is changed when stocks are added or removed from the index and when stocks split.

$$\text{DJIA} = \frac{\sum_{i=1}^{30} y_i}{C}$$

where y_i is the closing price for stock i and $C = 0.155625$. Using the stock prices given, compute the DJIA for June 19, 2014.

c. The DJIA is a summary of data. Does the DJIA provide information about a population using sampled data? If so, to what population? Is the sample a random sample?

Components of DJIA

Company	Stock Price (Noon 6/19/2014)
3M Co	144.41
American Express Co	94.72
AT&T Inc	35.31
Boeing Co	132.41
Caterpillar Inc	107.28
Chevron Corp	130.73
Cisco Systems Inc	24.63
E.I. Dupont de Nemours and Co	67.55
Exxon Mobil Corp	101.85
General Electric Co	26.90
Goldman Sachs Group Inc	169.52
Home Depot Inc	80.10
Intel Corp	29.99
IBM	182.95
Johnson & Johnson	103.41
JP Morgan Chase and Co	57.36
McDonald's Corp	l01.60
Merck & Co Inc	58.27
Microsoft Corp	41.45
Nike Inc	75.43
Pfizer	29.55
Procter & Gamble Co	80.28
The Coca-Cola Co	41.76
Travelers Companies Inc	95.51
United Technologies Corp	117.09
United Health Group lnc	79.88
Verizon Communications Inc	49.47
Visa Inc	208.30
Wal-Mart Stores Inc	76.25
Walt Disney Co	83.69

H.R. **3.59** As one part of a review of middle-manager selection procedures, a study was made of the relation between the hiring source (promoted from within, hired from related business, hired from unrelated business) and the 3-year job history (additional promotion, same position, resigned, dismissed). The data for 120 middle managers follow.

	Source			
Job History	Within Firm	Related Business	Unrelated Business	Total
Promoted	13	4	10	27
Same Position	32	8	18	58
Resigned	9	6	10	25
Dismissed	3	3	4	10
Total	57	21	42	120

a. Compute the job-history percentages within each of the three sources.
b. Describe the relation between job history and source.
c. Use an appropriate graph to display the relation between job history and source.

Env. **3.60** In order to assess the U.S. public's opinion about national energy policy, random samples were taken of 150 residents of major coal-producing states, 200 residents of major natural gas/oil-producing states, and 450 residents of the remaining states. Each resident was asked to select his or her most preferred national energy policy. The results are shown in the following table.

	Type of State			
Energy Policy	**Coal**	**Oil & Gas**	**Other**	**Total**
Coal Based	62	25	53	140
Oil & Gas Based	19	79	102	200
Nuclear Based	8	6	22	36
Solar & Wind Based	58	78	247	383
Fusion Based	3	12	26	41
Total	150	200	450	800

a. Replace the number of responses in the table with the five percentages for each of the three groups of respondents.
b. Based on the percentages, does there appear to be a strong dependence between the type of state and the energy policy?
c. Provide a graphical display of the dependency.
d. Which energy policy has the strongest support amongst the 800 surveyed people?
e. Do the opinions displayed in the above table represent the U.S. public's opinion in general?

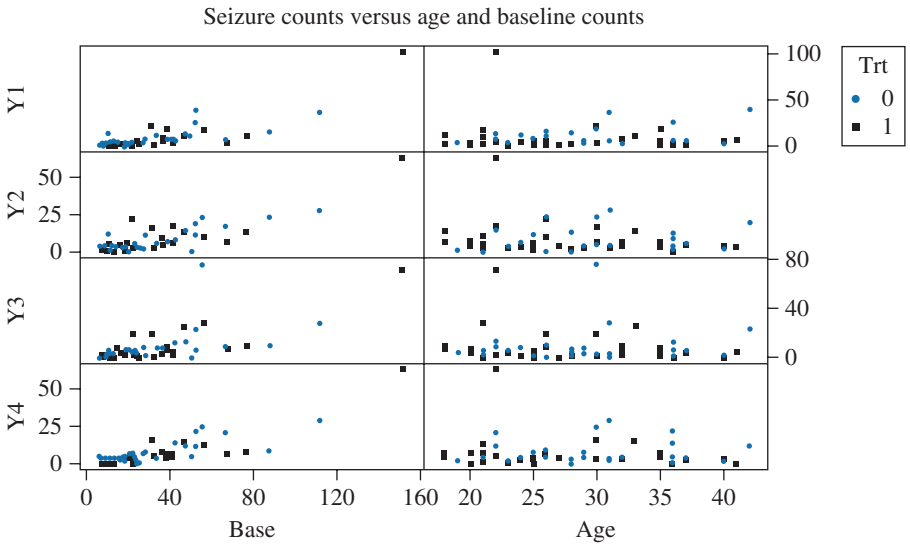
Bus. **3.61** A municipal workers' union that represents sanitation workers in many small midwestern cities studied the contracts that were signed in the previous years. The contracts were subdivided into those settled by negotiation without a strike, those settled by arbitration without a strike, and those settled after a strike. For each contract, the first-year percentage wage increase was determined. Summary figures follow.

Contract Type	**Negotation**	**Arbitration**	**Poststrike**
Mean Percentage Wage Increase	8.20	9.42	8.40
Variance	0.87	1.04	1.47
Standard Deviation	0.93	1.02	1.21
Sample Size	38	16	6

Does there appear to be a relationship between contract type and mean percentage wage increase? If you were management rather than union affiliated, which posture would you take in future contract negotiations?

Med. **3.62** Refer to the epilepsy study data in Table 3.19. Examine the scatterplots of Y_1, Y_2, Y_3, and Y_4 versus baseline count and age given here.

a. Does there appear to be a difference in the relationship between the seizure count ($Y_1 - Y_4$) and either the baseline count or age when considering the two groups (treatment and placebo)?
b. Describe the type of apparent differences, if any, that you found in part (a).



Med. **3.63** The correlations computed for the six variables in the epilepsy study are given here. Do the sizes of the correlation coefficients reflect the relationships displayed in the graphs given in Exercise 3.62? Explain your answer.

	Placebo Group				
	Y_1	Y_2	Y_3	Y_4	Base
Y_2	.782				
Y_3	.507	.661			
Y_4	.675	.780	.676		
Base	.744	.831	.493	.818	
Age	.326	.108	.113	.117	.033

	Treatment Group				
	Y_1	Y_2	Y_3	Y_4	Base
Y_2	.907				
Y_3	.912	.925			
Y_4	.971	.947	.952		
Base	.854	.845	.834	.876	
Age	−.141	−.243	−.194	−.197	−.343

Med. **3.64** An examination of the scatterplots in Exercise 3.62 reveals one patient with a very large value for baseline count and all subsequent counts. The patient has ID 207.

a. Predict the effect of removing the patient with ID 207 from the data set on the size of the correlations in the treatment group.

b. Using a computer program, compute the correlations with patient ID 207 removed from the data. Do the values confirm your predictions?

Med. **3.65** Refer to the research study concerning the effect of social factors on reading and math scores in Section 3.8. We justified studying just the reading scores because there was a strong correlation between reading and math scores. Construct the same plots for the math scores as were constructed for the reading scores.

a. Is there support for the same conclusions for the math scores as obtained for the reading scores?
b. If the conclusions are different, why do you suppose this has happened?

Med. **3.66** In the research study concerning the effect of social factors on reading and math scores, we found a strong negative correlation between %minority and %poverty and reading scores in Section 3.8.

a. Why is it not possible to conclude that large relative values for %minority and %poverty in a school result in lower reading scores for children in these social classes?
b. List several variables related to the teachers and students in the schools that may be important in explaining why low reading scores were strongly associated with schools having large values of %minority and %poverty.

Soc. **3.67** In the January 2004 issue of ***Consumer Reports,*** an article titled "Cut the Fat" described some of the possible problems in the diets of the U.S. public. The following table gives data on the increase in daily calories in the food supply per person. Construct a time-series plot to display the increase in calorie intake.

Year	1970	1975	1980	1985	1990	1995	2000
Calories	3,300	3,200	3,300	3,500	3,600	3,700	3,900

a. Describe the trend in calorie intake over the 30 years.
b. What would you predict the calorie intake was in 2005? Justify your answer by explaining any assumptions you are making about calorie intake.

Soc. **3.68** In the January 2004 issue of ***Consumer Reports,*** an article titled "Cut the Fat" described some of the possible problems in the diets of the U.S. public. The following table gives data on the increase in pounds of added sugar produced per person. Construct a time-series plot to display the increase in sugar production.

Year	1970	1975	1980	1985	1990	1995	2000
Pounds of Sugar	119	114	120	128	132	144	149

a. Describe the trend in sugar production over the 30 years.
b. Compute the correlation coefficient between calorie intake (using the data in Exercise 3.67) and sugar production. Is there strong evidence that the increase in sugar production is causing the increased calorie intake by the U.S. public?

Med. **3.69** Certain types of diseases tend to occur in clusters. In particular, persons affected with AIDS, syphilis, and tuberculosis may have some common characteristics and associations that increase their chances of contracting these diseases. The following table lists the number of reported cases by state in 2001.

State	AIDS	Syphilis	Tuber.	State	AIDS	Syphilis	Tuber.
AL	438	720	265	MT	15	0	20
AK	18	9	54	NE	74	16	40
AZ	540	1,147	289	NV	252	62	96
AR	199	239	162	NH	40	20	20
CA	4,315	3,050	3,332	NJ	1,756	1,040	530
CO	288	149	138	NM	143	73	54
CT	584	165	121	NY	7,476	3,604	1,676
DE	248	79	33	NC	942	1,422	398

(*continued*)

State	AIDS	Syphilis	Tuber.	State	AIDS	Syphilis	Tuber.
DC	870	459	74	ND	3	2	6
FL	5,138	2,914	1,145	OH	581	297	306
GA	1,745	1,985	575	OK	243	288	194
HI	124	41	151	OR	259	48	123
ID	19	11	9	PA	1,840	726	350
IL	1,323	1,541	707	RI	103	39	60
IN	378	529	115	SC	729	913	263
IA	90	44	43	SD	25	1	13
KS	98	88	63	TN	602	1,478	313
KY	333	191	152	TX	2,892	3,660	1,643
LA	861	793	294	UT	124	25	35
ME	48	16	20	VT	25	8	7
MD	1,860	937	262	VA	951	524	306
MA	765	446	270	WA	532	174	261
MI	548	1,147	330	WV	100	7	32
MN	157	132	239	WI	193	131	86
MS	418	653	154	WY	5	4	3
MO	445	174	157	All States	41,868	32,221	15,989

a. Construct a scatterplot of the number of AIDS cases versus the number of syphilis cases.
b. Compute the correlation between the number of AIDS cases and the number of syphilis cases.
c. Does the value of the correlation coefficient reflect the degree of association shown in the scatterplot?
d. Why do you think there may be a correlation between these two diseases?

Med. **3.70** Refer to the data in Exercise 3.69.
a. Construct a scatterplot of the number of AIDS cases versus the number of tuberculosis cases.
b. Compute the correlation between the number of AIDS cases and the number of tuberculosis cases.
c. Why do you think there may be a correlation between these two diseases?

Med. **3.71** Refer to the data in Exercise 3.69.
a. Construct a scatterplot of the number of syphilis cases versus the number of tuberculosis cases.
b. Compute the correlation between the number of syphilis cases and the number of tuberculosis cases.
c. Why do you think there may be a correlation between these two diseases?

Med. **3.72** Refer to the data in Exercise 3.69.
a. Construct a quantile plot of the number of syphilis cases.
b. From the quantile plot, determine the 90th percentile for the number of syphilis cases.
c. Identify the states in which the number of syphilis cases is above the 90th percentile.

Med. **3.73** Refer to the data in Exercise 3.69.
a. Construct a quantile plot of the number of tuberculosis cases.
b. From the quantile plot, determine the 90th percentile for the number of tuberculosis cases.
c. Identify the states in which the number of tuberculosis cases is above the 90th percentile.

Med. **3.74** Refer to the data in Exercise 3.69.

- **a.** Construct a quantile plot of the number of AIDS cases.
- **b.** From the quantile plot, determine the 90th percentile for the number of AIDS cases.
- **c.** Identify the states in which the number of AIDS cases is above the 90th percentile.

Med. **3.75** Refer to the results from Exercises 3.72–3.74.

- **a.** How many states had numbers of AIDS, tuberculosis, and syphilis cases that were all above the 90th percentiles?
- **b.** Identify these states and comment on any common elements among the states.
- **c.** How could the U.S. government apply the results from Exercises 3.69–3.75 in making public health policy?

Med. **3.76** The article ***"Viral Load and Heterosexual Transmission of Human Immunodeficiency Virus Type 1"*** **[New England Journal of Medicine *(2000) 342:921–929]***, reports a study that addressed the question of whether people with high levels of HIV-1 are significantly more likely to transmit HIV to their uninfected partners. Measurements follow of the HIV-1 RNA levels in the group whose partners were initially uninfected but became HIV positive during the course of the study: values are given in units of RNA copies/mL.

79725, 12862, 18022, 76712, 256440, 14013, 46083, 6808, 85781, 1251,
6081, 50397, 11020, 13633 1064, 496433, 25308, 6616, 11210, 13900

- **a.** Determine the mean, median, and standard deviation.
- **b.** Find the 25th, 50th, and 75th percentiles.
- **c.** Plot the data in a boxplot and histogram.
- **d.** Describe the shape of the distribution.

Med. **3.77** In many statistical procedures, it is often advantageous to have a symmetric distribution. When the data have a histogram that is highly right-skewed, it is often possible to obtain a symmetric distribution by taking a transformation of the data. For the data in Exercise 3.76, take the natural logarithm of the data and answer the following questions.

- **a.** Determine the mean, median, and standard deviation.
- **b.** Find the 25th, 50th, and 75th percentiles.
- **c.** Plot the data in a boxplot and histogram.
- **d.** Did the logarithm transformation result in a somewhat symmetric distribution?

Env. **3.78** PCBs are a class of chemicals often found near the disposal of electrical devices. PCBs tend to concentrate in human fat and have been associated with numerous health problems. In the article ***"Some Other Persistent Organochlorines in Japanese Human Adipose Tissue"*** **[Environmental Health Perspective (April, 2000) *108:599—603]***, researchers examined the concentrations of PCB (ng/g) in the fat of a group of adults. They detected the following concentrations:

1800, 1800, 2600, 1300, 520, 3200, 1700, 2500, 560, 930, 2300, 2300, 1700, 720

- **a.** Determine the mean, median, and standard deviation.
- **b.** Find the 25th, 50th, and 75th percentiles.
- **c.** Plot the data in a boxplot.
- **d.** Would it be appropriate to apply the Empirical Rule to these data? Why or why not?

Ag. **3.79** The focal point of an agricultural research study was the relationship between the time a crop is planted and the amount of crop harvested. If a crop is planted too early or too late, farmers may fail to obtain optimal yield and hence not make a profit. An ideal date for planting is set by the researchers, and the farmers then record the number of days either before or after the designated date. In the following data set, D is the number of days from the ideal planting date and Y is the yield (in bushels per acre) of a wheat crop:

D	−19	−18	−15	−12	−9	−6	−4	−3	−1	0
Y	30.7	29.7	44.8	41.4	48.1	42.8	49.9	46.9	46.4	53.5
D	1	3	6	8	12	15	17	19	21	24
Y	55.0	46.9	44.1	50.2	41.0	42.8	36.5	35.8	32.2	23.3

a. Plot the data in a scatterplot.
b. Describe the relationship between the number of days from the optimal planting date and the wheat yield.
c. Calculate the correlation coefficient between days from optimal planting and yield.
d. Explain why the correlation coefficient is relatively small for this data set.

Con. **3.80** Although an exhaust fan is present in nearly every bathroom, it often is not used due to the high noise level. This is an unfortunate practice because regular use of the fan results in a reduction of indoor moisture. Excessive indoor moisture often results in the development of mold, which may have adverse health consequences. ***Consumer Reports*** in its January 2004 issue reports on a wide variety of bathroom fans. The following table displays the price (P) in dollars of the fans and the quality of the fan measured in airflow (AF), cubic feet per minute (cfm).

P	95	115	110	15	20	20	75	150	60	60
AF	60	60	60	55	55	55	85	80	80	75
P	160	125	125	110	130	125	30	60	110	85
AF	90	90	100	110	90	90	90	110	110	60

a. Plot the data in a scatterplot and comment on the relationship between price and airflow.
b. Compute the correlation coefficient for this data set. Is there a strong or weak relationship between price and airflow of the fans?
c. Is your conclusion in part (b) consistent with your answer in part (a)?
d. Based on your answers in parts (a) and (b), would it be reasonable to conclude that higher-priced fans generate greater airflow?

CHAPTER 4

Probability and Probability Distributions

4.1 Introduction and Abstract of Research Study

We stated in Chapter 1 that a scientist uses inferential statistics to make statements about a population based on information contained in a sample of units selected from that population. Graphical and numerical descriptive techniques were presented in Chapter 3 as a means to summarize and describe a sample.

However, a sample is not identical to the population from which it was selected. We need to assess the degree of accuracy to which the sample mean, sample standard deviation, or sample proportion represents the corresponding population values.

Most management decisions must be made in the presence of uncertainty. Prices and designs for new automobiles must be selected on the basis of shaky forecasts of consumer preference, national economic trends, and competitive actions. The size and allocation of a hospital staff must be decided with limited information on patient load. The inventory of a product must be set in the face of uncertainty about demand. Probability is the language of uncertainty. Now let us examine probability, the mechanism for making inferences. This idea is probably best illustrated by an example.

Newsweek, in its ***June 20, 1998,*** issue, asks the question "Who Needs Doctors? The Boom in Home Testing." The article discusses the dramatic increase in medical screening tests for home use. The home-testing market has expanded beyond the two most frequently used tests, pregnancy and diabetes glucose monitoring, to a variety of diagnostic tests that were previously used only by doctors and certified laboratories. There is a DNA test to determine whether twins are fraternal or identical, a test to check cholesterol level, a screening test for colon cancer, and tests to determine whether your teenager is a drug user. However, the major question that needs to be addressed is, How reliable are the testing kits? When a test indicates that a woman is not pregnant, what is the chance that the test is incorrect and the woman is truly pregnant? This type of incorrect result from a home test could translate into a woman not seeking the proper prenatal care in the early stages of her pregnancy.

Suppose a company states in its promotional materials that its pregnancy test provides correct results in 75% of its applications by pregnant women. We want to evaluate the claim, so we select 20 women who have been determined by their physicians, using the best possible testing procedures, to be pregnant. The test is taken by each of the 20 women, and for all 20 women, the test result is negative, indicating that none of the 20 is pregnant. What do you conclude about the company's claim about the reliability of its test? Suppose you are further assured that each of the 20 women was in fact pregnant, as was determined several months after the test was taken.

If the company's claim of 75% reliability was correct, we would have expected somewhere near 75% of the tests in the sample to be positive. However, none of the test results was positive. Thus, we would conclude that the company's claim is probably false. Why did we fail to state with certainty that the company's claim was false? Consider the possible setting. Suppose we have a large population consisting of millions of units and 75% of the units are Ps for positives and 25% of the units are Ns for negatives. We randomly select 20 units from the population and count the number of units in the sample that are Ps. Is it possible to obtain a sample consisting of 0 Ps and 20 Ns? Yes, it is possible, *but* it is highly *improbable*. Later in this chapter, we will compute the probability of such a sample occurrence.

To obtain a better view of the role that probability plays in making inferences from sample results that are then used to draw conclusions about populations, suppose 14 of the 20 tests are positive—that is, a 70% correct response rate. Would you consider this result highly improbable and reject the company's claim of a 75% correct response rate? How about 12 positives and 8 negatives, or 16 positives and 4 negatives? At what point do we decide that the result of the

observed sample is so improbable, assuming the company's claim is correct, that we disagree with its claim? To answer this question, we must know how to find the probability of obtaining a particular sample outcome. Knowing this probability, we can then determine whether we agree or disagree with the company's claim. Probability is the tool that enables us to make an inference. Later in this chapter, we will discuss in detail how the FDA and private companies determine the reliability of screening tests.

Because probability is the tool for making inferences, we need to define probability. In the preceding discussion, we used the term *probability* in its everyday sense. Let us examine this idea more closely.

Observations of phenomena can result in many different outcomes, some of which are more likely than others. Numerous attempts have been made to give a precise definition for the probability of an outcome. We will cite three of these.

classical interpretation of probability

The first interpretation of probability, called the **classical interpretation of probability,** arose from games of chance. Typical probability statements of this type are, for example, "the probability that a flip of a balanced coin will show 'heads' is 1/2" and "the probability of drawing an ace when a single card is drawn from a standard deck of 52 cards is 4/52." The numerical values for these probabilities arise from the nature of the games. A coin flip has two possible outcomes (a head or a tail); the probability of a head should then be 1/2 (1 out of 2). Similarly, there are 4 aces in a standard deck of 52 cards, so the probability of drawing an ace in a single draw is 4/52, or 4 out of 52.

outcome
event

In the classical interpretation of probability, each possible distinct result is called an **outcome;** an **event** is identified as a collection of outcomes. The probability of an event E under the classical interpretation of probability is computed by taking the ratio of the number of outcomes, N_e, favorable to event E to the total number of possible outcomes, N:

$$P(\text{event } E) = \frac{N_e}{N}$$

The applicability of this interpretation depends on the assumption that all outcomes are equally likely. If this assumption does not hold, the probabilities indicated by the classical interpretation of probability will be in error.

relative frequency interpretation

A second interpretation of probability is called the **relative frequency concept of probability;** this is an empirical approach to probability. If an experiment is repeated a large number of times and event E occurs 30% of the time, then .30 should be a very good approximation to the probability of event E. Symbolically, if an experiment is conducted n different times and if event E occurs on n_e of these trials, then the probability of event E is approximately

$$P(\text{event } E) \cong \frac{n_e}{n}$$

We say "approximately" because we think of the actual probability $P(\text{event } E)$ as the relative frequency of the occurrence of event E over a very large number of observations or repetitions of the phenomenon. The fact that we can check probabilities that have a relative frequency interpretation (by simulating many repetitions of the experiment) makes this interpretation very appealing and practical.

The third interpretation of probability can be used for problems in which it is difficult to imagine a repetition of an experiment. These are "one-shot" situations. For example, the director of a state welfare agency who estimates the probability that a proposed revision in eligibility rules will be passed by the state legislature

personal/subjective interpretation of probability

would not be thinking in terms of a long series of trials. Rather, the director would use a **personal** or **subjective probability** to make a one-shot statement of belief regarding the likelihood of passage of the proposed legislative revision. The problem with subjective probabilities is that they can vary from person to person and they cannot be checked.

Of the three interpretations presented, the relative frequency concept seems to be the most reasonable one because it provides a practical interpretation of the probability for most events of interest. Even though we will never run the necessary repetitions of the experiment to determine the exact probability of an event, the fact that we can check the probability of an event gives meaning to the relative frequency concept. Throughout the remainder of this text, we will lean heavily on this interpretation of probability.

Abstract of Research Study: Inferences About Performance-Enhancing Drugs Among Athletes

The ***Associated Press*** reported the following in an April 28, 2005, article:

> CHICAGO—The NBA and its players union are discussing expanded testing for performance-enhancing drugs, and commissioner David Stern said Wednesday he is optimistic it will be part of the new labor agreement. The league already tests for recreational drugs and more than a dozen types of steroids. But with steroid use by professional athletes and the impact they have on children under increasing scrutiny, Stern said he believes the NBA should do more.

An article in ***USA Today (April 27, 2005)*** by Dick Patrick reports:

> Just before the House Committee on Government Reform hearing on steroids and the NFL ended Wednesday, ranking minority member Henry Waxman, D-Calif., expressed his ambiguity about the effectiveness of the NFL testing system. He spoke to a witness panel that included NFL Commissioner Paul Tagliabue and NFL Players Association executive director Gene Upshaw, both of whom had praised the NFL system and indicated there was no performance-enhancing drug problem in the league. "There's still one thing that puzzles me," Waxman said, "and that's the fact that there are a lot of people who are very credible in sports who tell me privately that there's a high amount of steroid use in football. When I look at the testing results, it doesn't appear that's the case. It's still nagging at me."

Finally, we have a report from ***ABC News (April 27, 2005)*** in which the drug issue in major league sports is discussed:

> A law setting uniform drug-testing rules for major U.S. sports would be a mistake, National Football League Commissioner Paul Tagliabue said Wednesday under questioning from House lawmakers skeptical that professional leagues are doing enough. "We don't feel that there is rampant cheating in our sport," Tagliabue told the House Government Reform Committee. Committee members were far less adversarial than they were last month, when Mark McGwire, Jose Canseco and other current and former baseball stars were compelled to appear and faced tough questions about steroid use. Baseball commissioner Bud Selig, who also appeared at that hearing, was roundly criticized for the punishments in his sport's policy, which lawmakers said was too lenient.

One of the major reasons the leaders of professional sports athletes' unions are so concerned about drug testing is that failing a drug test can devastate an athlete's career. The controversy over performance-enhancing drugs has seriously brought into question the reliability of the tests for these drugs. Some banned substances, such as stimulants like cocaine and artificial steroids, are relatively easy to

deal with because they are not found naturally in the body. If these are detected at all, the athlete is banned. Nandrolone, a close chemical cousin of testosterone, was thought to be in this category until recently. But a study has since shown that normal people can have a small but significant level in their bodies—0.6 nanograms per milliliter of urine. The International Olympic Committee has set a limit of 2 nanograms per milliliter. But expert Mike Wheeler, a doctor at St Thomas' Hospital, states that this is "awfully close" to the level at which an unacceptable number (usually more than .01%) of innocent athletes might produce positive tests.

In an article titled "Inferences About Testosterone Abuse Among Athletes," in a 2004 issue of ***Chance (17:5–8)***, the authors discuss some of the issues involved with the drug testing of athletes. In particular, they discuss the issues involved in determining the reliability of drug tests. They report:

> The diagnostic accuracy of any laboratory test is defined as the ability to discriminate between two types of individuals—in this case, users and nonusers. *Specificity* and *sensitivity* characterize diagnostic tests.... Estimating these proportions requires collecting and tabulating data from the two reference samples, users and nonusers,... Bayes' rule is a necessary tool for relating experimental evidence to conclusions, such as whether someone has a disease or has used a particular substance. Applying Bayes' rule requires determining the test's sensitivity and specificity. It also requires a pre-test (or prior) probability that the athlete has used a banned substance.

Any drug test can result in a false positive due to the variability in the testing procedure, biologic variability, or inadequate handling of the material to be tested. Even if a test is highly reliable and produces only 1% false positives but the test is widely used, with 80,000 tests run annually, the result would be that 800 athletes would be falsely identified as using a banned substance. The result is that innocent people will be punished. The trade-off between determining that an athlete is a drug user and convincing the public that the sport is being conducted fairly is not obvious. The authors state, "Drug testing of athletes has two purposes: to prevent artificial performance enhancement (known as doping) and to discourage the use of potentially harmful substances." Thus, there is a need to be able to assess the reliability of any testing procedure.

In this chapter, we will explicitly define the terms *specificity, sensitivity,* and *prior probability*. We will then formulate *Bayes' rule* (which we will designate as Bayes' Formula). At the end of the chapter, we will return to this article and discuss the issues of *false positives* and *false negatives* in drug testing and how they are computed from our knowledge of the specificity and sensitivity of a drug test along with the prior probability that a person is a user.

4.2 Finding the Probability of an Event

In the preceding section, we discussed three different interpretations of probability. In this section, we will use the classical interpretation and the relative frequency concept to illustrate the computation of the probability of an outcome or event. Consider an experiment that consists of tossing two coins, a penny and then a dime, and observing the upturned faces. There are four possible outcomes:

TT: tails for both coins
TH: a tail for the penny, a head for the dime
HT: a head for the penny, a tail for the dime
HH: heads for both coins

What is the probability of observing the event exactly one head from the two coins?

This probability can be obtained easily if we can assume that all four outcomes are equally likely. In this case, that seems quite reasonable. There are $N = 4$ possible outcomes, and $N_e = 2$ of these are favorable for the event of interest, observing exactly one head. Hence, by the classical interpretation of probability,

$$P(\text{exactly 1 head}) = \frac{2}{4} = \frac{1}{2}$$

Because the event of interest has a relative frequency interpretation, we could also obtain this same result empirically, using the relative frequency concept. To demonstrate how relative frequency can be used to obtain the probability of an event, we will use the ideas of simulation. Simulation is a technique that produces outcomes having the same probability of occurrence as the real situation events. The computer is a convenient tool for generating these outcomes. Suppose we wanted to simulate 500 tosses of two fair coins. We can use a computer program R to simulate the tosses. R is a software program that you can obtain free of charge by visiting the website *cran.r-product.org* or just by typing **CRAN** into Google. The following R code will be used to generate 500 two-digit numbers. Even digits will be designated as H and odd digits designated as T. The 500 numbers have now been transformed into pairs of Ts and Hs. Because there are five even and five odd single-digit numbers, the probability of obtaining an even number is 5/10 = .5, which is the same probability of obtaining an odd number. This set of 500 pairs of single-digit numbers represents 500 tosses of two fair coins; that is, coins in which the probabilities of H and T are both .5. The first digit represents the outcome of tossing the first coin and, the second digit represents the toss of the second coin. For example, the number 36 would represent a T for the toss of the first coin and an H for the toss of the second coin. The following lines of code in R will generate 500 pairs of randomly selected single-digit numbers.

1. y = c(0:9)
2. x_1 = sample(y, 500, replace = T)
3. x_2 = sample(y, 500, replace = T)
4. x = cbind(x_1, x_2)
5. x

Most computer packages contain a random-number generator that can be used to produce similar results. Table 4.1(a) contains the results of the simulation of the 500 pairs of tosses. The 500 pairs of single-digit numbers are then summarized in Table 4.1(b).

Note that this approach yields simulated probabilities that are nearly in agreement with our intuition; that is, intuitively we might expect these outcomes to be equally likely. Thus, each of the four outcomes should occur with a probability equal to 1/4, or .25. This assumption was made for the classical interpretation. We will show in Chapter 10 that in order to be 95% certain that the simulated probabilities are within .01 of the true probabilities, the number of tosses should be at least 7,500 and not 500 as we used previously.

If we wish to find the probability of tossing two coins and observing exactly one head, we have, from Table 4.1(b),

$$P(\text{exactly 1 head}) \cong \frac{117 + 125}{500} = .484$$

TABLE 4.1(a) Simulation of tossing a penny and a dime 500 times

25	32	70	15	96	87	80	43	15	77	89	51	08	36	29	55	42	86	45	93	68	72	49	99	37
82	81	58	50	85	27	99	41	10	31	42	35	50	02	68	33	50	93	73	62	15	15	90	97	24
46	86	89	82	20	23	63	59	50	40	32	72	59	62	58	53	01	85	49	27	31	48	53	07	78
15	81	39	83	79	21	88	57	35	33	49	37	85	42	28	38	50	43	82	47	01	55	42	02	52
66	44	15	40	29	73	11	06	79	81	49	64	32	06	07	31	07	78	73	07	26	36	39	20	14
48	20	27	73	53	21	44	16	00	33	43	95	21	08	19	60	68	30	99	27	22	74	65	22	05
26	79	54	64	94	01	21	47	86	94	24	41	06	81	16	07	30	34	99	54	68	37	38	71	79
86	12	83	09	27	60	49	54	21	92	64	57	07	39	04	66	73	76	74	93	50	56	23	41	23
18	87	21	48	75	63	09	97	96	86	85	68	65	35	92	40	57	87	82	71	04	16	01	03	45
52	79	14	12	94	51	39	40	42	17	32	94	42	34	68	17	39	32	38	03	75	56	79	79	57
07	40	96	46	22	04	12	90	80	71	46	11	18	81	54	95	47	72	06	07	66	05	59	34	81
66	79	83	82	62	20	75	71	73	79	48	86	83	74	04	13	36	87	96	11	39	81	59	41	70
21	47	34	02	05	73	71	57	64	58	05	16	57	27	66	92	97	68	18	52	09	45	34	80	57
87	22	18	65	66	18	84	31	09	38	05	67	10	45	03	48	52	48	33	36	00	49	39	55	35
70	84	50	37	58	41	08	62	42	64	02	29	33	68	87	58	52	39	98	78	72	13	13	15	96
57	32	98	05	83	39	13	39	37	08	17	01	35	13	98	66	89	40	29	47	37	65	86	73	42
85	65	78	05	24	65	24	92	03	46	67	48	90	60	02	61	21	12	80	70	35	15	40	52	76
29	11	45	22	38	33	32	52	17	20	03	26	34	18	85	46	52	66	63	30	84	53	76	47	21
42	97	56	38	41	87	14	43	30	35	99	06	76	67	00	47	83	32	52	42	48	51	69	15	18
08	30	37	89	17	89	23	58	13	93	17	44	09	08	61	05	35	44	91	89	35	15	06	39	27

TABLE 4.1(b)
Summary of the simulation

Event	Outcome of Simulation	Frequency	Relative Frequency
TT	(Odd, Odd)	129	129/500 = .258
TH	(Odd, Even)	117	117/500 = .234
HT	(Even, Odd)	125	125/500 = .250
HH	(Even, Even)	129	129/500 = .258

This is very close to the theoretical probability, which we have shown to be .5.

Note that we could easily modify our example to accommodate the tossing of an unfair coin. Suppose we are tossing a penny that is weighted so that the probability of a head occurring in a toss is .70 and the probability of a tail is .30. We could designate an H outcome whenever one of the random digits 0, 1, 2, 3, 4, 5, or 6 occurs and a T outcome whenever one of the digits 7, 8, or 9 occurs. The same simulation program can be run as before, but we would interpret the output differently.

4.3 Basic Event Relations and Probability Laws

The probability of an event—say, event A—will always satisfy the property

$$0 \leq P(A) \leq 1$$

that is, the probability of an event lies anywhere in the interval from 0 (the occurrence of the event is impossible) to 1 (the occurrence of the event is a "sure thing").

either *A* or *B* occurs

Suppose A and B represent two experimental events and you are interested in a new event, the event that **either *A* or *B* occurs.** For example, suppose that we toss a pair of dice and define the following events:

A: A total of 7 shows
B: A total of 11 shows

Then the event "either A or B occurs" is the event that you toss a total of either 7 or 11 with the pair of dice.

mutually exclusive

Note that, for this example, the events A and B are **mutually exclusive;** that is, if you observe event A (a total of 7), you could not at the same time observe event B (a total of 11). Thus, if A occurs, B cannot occur (and vice versa).

DEFINITION 4.1

Two events A and B are said to be **mutually exclusive** if (when the experiment is performed a single time) the occurrence of one of the events excludes the possibility of the occurrence of the other event.

The concept of mutually exclusive events is used to specify a second property that the probabilities of events must satisfy. When two events are mutually exclusive, then the probability that either one of the events will occur is the sum of the event probabilities.

DEFINITION 4.2

If two events, A and B, are mutually exclusive, the **probability** that either event occurs is $P(\text{either } A \text{ or } B) = P(A) + P(B)$.

Definition 4.2 is a special case of the union of two events, which we will soon define.

The definition of additivity of probabilities for mutually exclusive events can be extended beyond two events. For example, when we toss a pair of dice, the sum S of the numbers appearing on the dice can assume any one of the values $S = 2, 3, 4, \ldots, 11, 12$. On a single toss of the dice, we can observe only one of these values. Therefore, the values $2, 3, \ldots, 12$ represent mutually exclusive events. If we want to find the probability of tossing a sum less than or equal to 4, this probability is

$$P(S \leq 4) = P(2) + P(3) + P(4)$$

For this particular experiment, the dice can fall in 36 different equally likely ways. We can observe a 1 on die 1 and a 1 on die 2, denoted by the symbol (1, 1). We can observe a 1 on die 1 and a 2 on die 2, denoted by (1, 2). In other words, for this experiment, the possible outcomes are

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)
(1, 3)	(2, 3)	(3, 3)	(4, 3)	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	(4, 4)	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	(4, 5)	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	(6, 6)

As you can see, only one of these events, (1, 1), will result in a sum equal to 2. Therefore, we would expect a 2 to occur with a relative frequency of 1/36 in a long series of repetitions of the experiment, and we let $P(2) = 1/36$. The sum $S = 3$ will occur if we observe an outcome of either (1, 2) or (2, 1). Therefore, $P(3) = 2/36 = 1/18$. Similarly, we find $P(4) = 3/36 = 1/12$. It follows that

$$P(S \leq 4) = P(2) + P(3) + P(4) = \frac{1}{36} + \frac{1}{18} + \frac{1}{12} = \frac{1}{6}$$

complement

A third property of event probabilities concerns an event and its **complement.**

DEFINITION 4.3

The **complement** of an event A is the event that A *does not* occur. The complement of A is denoted by the symbol $\overline{A}$.

Thus, if we define the complement of an event A as a new event—namely, "A does not occur"—it follows that

$$P(A) + P(\overline{A}) = 1$$

For an example, refer again to the two-coin-toss experiment. If, in many repetitions of the experiment, the proportion of times you observe event A, "two heads show," is $1/4$, then it follows that the proportion of times you observe the event $\overline{A}$, "two heads do not show," is $3/4$. Thus, $P(A)$ and $P(\overline{A})$ will always sum to 1.

We can summarize the three properties that the probabilities of events must satisfy as follows:

Properties of Probabilities

If A and B are any two mutually exclusive events associated with an experiment, then $P(A)$ and $P(B)$ must satisfy the following properties:

1. $0 \le P(A) \le 1$ and $0 \le P(B) \le 1$
2. $P(\text{either } A \text{ or } B) = P(A) + P(B)$
3. $P(A) + P(\overline{A}) = 1$ and $P(B) + P(\overline{B}) = 1$

union
intersection

We can now define two additional event relations: the **union** and the **intersection** of two events.

DEFINITION 4.4

The **union** of two events A and B is the set of all outcomes that are included in either A or B (or both). The union is denoted as $A \cup B$.

DEFINITION 4.5

The **intersection** of two events A and B is the set of all outcomes that are included in both A and B. The intersection is denoted as $A \cap B$.

These definitions along with the definition of the complement of an event formalize some simple concepts. The event $\overline{A}$ occurs when A *does not;* $A \cup B$ occurs when either A or B occurs; $A \cap B$ occurs when A *and* B occur.

The additivity of probabilities for mutually exclusive events, called the *addition law for mutually exclusive events,* can be extended to give the general addition law.

DEFINITION 4.6

Consider two events A and B; the **probability of the union** of A and B is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

EXAMPLE 4.1

Events and event probabilities are shown in the Venn diagram in Figure 4.1. Use this diagram to determine the following probabilities:

a. $P(A), P(\overline{A})$
b. $P(B), P(\overline{B})$
c. $P(A \cap B)$
d. $P(A \cup B)$

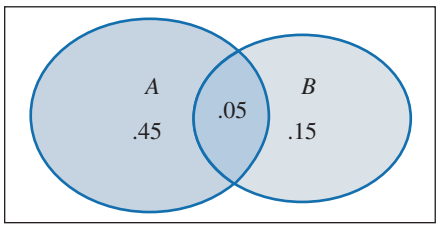


FIGURE 4.1
Probabilities for events A and B

Solution From the Venn diagram, we are able to determine the following probabilities:

a. $P(A) = .5$; therefore $P(\overline{A}) = 1 - .5 = .5$
b. $P(B) = .2$; therefore $P(\overline{B}) = 1 - .2 = .8$
c. $P(A \cap B) = .05$
d. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = .5 + .2 - .05 = .65$ ■

4.4 Conditional Probability and Independence

Consider the following situation: The examination of a large number of insurance claims, categorized according to type of insurance and whether the claim was fraudulent, produced the results shown in Table 4.2. Suppose you are responsible for checking insurance claims—in particular, for detecting fraudulent claims—and you examine the next claim that is processed. What is the probability of the event F, "the claim is fraudulent"? To answer the question, you examine Table 4.2 and note that 10% of all claims are fraudulent. Thus, assuming that the percentages given in the table are reasonable approximations to the true probabilities of receiving specific types of claims, it follows that $P(F) = .10$. Would you say that the risk that you face a fraudulent claim has probability .10? We think not, because you have additional information that may affect the assessment of $P(F)$. This additional information concerns the type of policy you are examining (fire, auto, or other).

TABLE 4.2
Categorization of insurance claims

	Type of Policy (%)			
Category	**Fire**	**Auto**	**Other**	**Total %**
Fraudulent	6	1	3	10
Nonfraudulent	14	29	47	90
Total	20	30	50	100

Suppose that you have the additional information that the claim was associated with a fire policy. Checking Table 4.2, we see that 20% (or .20) of all claims are associated with a fire policy and that 6% (or .06) of all claims are fraudulent fire policy claims. Therefore, it follows that the probability that the claim is fraudulent, given that you know the policy is a fire policy, is

$$P(F|\text{fire policy}) = \frac{\text{proportion of claims that are fraudulent fire policy claims}}{\text{proportion of claims that are against fire policies}}$$

$$= \frac{.06}{.20} = .30$$

conditional probability

This probability, $P(F|\text{fire policy})$, is called a **conditional probability** of the event F—that is, the probability of event F given the fact that the event "fire policy" has already occurred. This tells you that 30% of all fire policy claims are fraudulent. The vertical bar in the expression $P(F|\text{fire policy})$ represents the phrase "given that," or simply "given." Thus, the expression is read, "the probability of the event F given the event fire policy."

unconditional/ marginal probability

The probability $P(F) = .10$, called the **unconditional** or **marginal probability** of the event F, gives the proportion of times a claim is fraudulent—that is, the proportion of times event F occurs in a very large (infinitely large) number of repetitions of the experiment (receiving an insurance claim and determining whether the claim is fraudulent). In contrast, the conditional probability of F, given that the claim is for a fire policy, $P(F|\text{fire policy})$, gives the proportion of fire policy claims that are fraudulent. Clearly, the conditional probabilities of F, given the types of policies, will be of much greater assistance in measuring the risk of fraud than the unconditional probability of F.

DEFINITION 4.7

Consider two events A and B with nonzero probabilities, $P(A)$ and $P(B)$. The **conditional probability** of event A, given event B, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The conditional probability of event B, given event A, is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

This definition for conditional probabilities gives rise to what is referred to as the *multiplication law*.

DEFINITION 4.8

The **probability of the intersection** of two events A and B is

$$P(A \cap B) = P(A)P(B|A)$$
$$= P(B)P(A|B)$$

The only difference between Definitions 4.7 and 4.8, both of which involve conditional probabilities, relates to what probabilities are known and what needs to be calculated. When we know the intersection probability $P(A \cap B)$ and the individual probability $P(A)$, we can compute $P(B|A)$. When we know $P(A)$ and $P(B|A)$, we can compute $P(A \cap B)$.

EXAMPLE 4.2

A corporation is proposing to select 2 of its current regional managers as vice presidents. In the history of the company, there has never been a female vice president. The corporation has 6 male regional managers and 4 female regional managers. Make the assumption that the 10 regional managers are equally qualified and hence all possible groups of 2 managers should have the same chance of being selected as the vice presidents. Now find the probability that both vice presidents are male.

Solution Let A be the event that the first vice president selected is male, and let B be the event that the second vice president selected is also male. The event that represents both selected vice presidents are male is the event (A and B)—that is, the event $A \cap B$. Therefore, we want to calculate $P(A \cap B) = P(B|A)P(A)$, using Definition 4.8.

For this example,

$$P(A) = P(\text{first selection is male}) = \frac{\#\text{ of male managers}}{\#\text{ of managers}} = \frac{6}{10}$$

and

$$P(B|A) = P(\text{second selection is male, given first selection was male})$$
$$= \frac{\#\text{ of male managers after one male manager was selected}}{\#\text{ of managers after one male manager was selected}} = \frac{5}{9}$$

Thus,

$$P(A \cap B) = P(A)P(B|A) = \frac{6}{10}\left(\frac{5}{9}\right) = \frac{30}{90} = \frac{1}{3}$$

Thus, the probability that both vice presidents are male is 1/3 under the condition that all candidates are equally qualified and that each group of two managers has the same chance of being selected. Thus, there is a relatively large probability of selecting two males as the vice presidents under the condition that all candidates are equally likely to be selected. ■

Suppose that the probability of event A is the same whether event B has or has not occurred; that is, suppose

$$P(A|B) = P(A|\overline{B}) = P(A)$$

independent events

dependent events

Then we say that the occurrence of event A is not dependent on the occurrence of event B, or simply that A and B are **independent events.** When $P(A|B) \neq P(A)$, the occurrence of A depends on the occurrence of B, and events A and B are said to be **dependent events.**

DEFINITION 4.9

Two events A and B are **independent events** if

$$P(A|B) = P(A) \text{ or } P(B|A) = P(B)$$

(*Note:* You can show that if $P(A|B) = P(A)$, then $P(B|A) = P(B)$, and vice versa.)

Definition 4.9 leads to a special case of $P(A \cap B)$. When events A and B are independent, it follows that

$$P(A \cap B) = P(A)P(B|A) = P(A)P(B)$$

The concept of independence is of particular importance in sampling. Later in the text, we will discuss drawing samples from two (or more) populations to compare the means, variances, or other population parameters. For most of these applications, we will select samples in such a way that the observed values in one sample are independent of the values that appear in another sample. We call these **independent samples.**

independent samples

4.5 Bayes' Formula

In this section, we will show how Bayes' Formula can be used to update conditional probabilities by using sample data when available. These "updated" conditional probabilities are useful in decision making. A particular application of these techniques involves the evaluation of diagnostic tests. Suppose a meat inspector must decide whether a randomly selected meat sample contains *E. coli* bacteria. The inspector conducts a diagnostic test. Ideally, a positive result (Pos) would mean that the meat sample actually has *E. coli,* and a negative result (Neg) would imply that the meat sample is free of *E. coli.* However, the diagnostic test is occasionally in error. The result of the test may be a **false positive,** for which the test's indication of *E. coli* presence is incorrect, or a **false negative,** for which the test's conclusion of *E. coli* absence is incorrect. Large-scale screening tests are conducted to evaluate the accuracy of a given diagnostic test. For example, *E. coli* (E) is placed in 10,000 meat samples, and the diagnostic test yields a positive result for 9,500 samples and a negative result for 500 samples; that is, there are 500 false negatives out of the 10,000 tests. Another 10,000 samples have all traces of *E. coli* removed (indicated as NE), and the diagnostic test yields a positive result for 100 samples and a negative result for 9,900 samples. There are 100 false positives out of the 10,000 tests. We can summarize the results in Table 4.3.

false positive
false negative

Evaluation of test results is as follows:

$$\text{True positive rate} = P(\text{Pos}|\text{E}) = \frac{9{,}500}{10{,}000} = .95$$

$$\text{False positive rate} = P(\text{Pos}|\text{NE}) = \frac{100}{10{,}000} = .01$$

$$\text{True negative rate} = P(\text{Neg}|\text{NE}) = \frac{9{,}900}{10{,}000} = .99$$

$$\text{False negative rate} = P(\text{Neg}|\text{E}) = \frac{500}{10{,}000} = .05$$

TABLE 4.3
E. coli test data

Diagnostic Test Result	Meat Sample Status	
	E	NE
Positive	9,500	100
Negative	500	9,900
Total	10,000	10,000

The **sensitivity** of the diagnostic test is the true positive rate—that is, P(test is positive|disease is present). The **specificity** of the diagnostic test is the true negative rate—that is, P(test is negative|disease is not present).

sensitivity
specificity

The primary task facing the inspector is to evaluate the probability of *E. coli* being present in the meat sample when the test yields a positive result—that is, the inspector needs to know $P(\text{E}|\text{Pos})$. Bayes' Formula provides us with a method to obtain this probability.

Bayes' Formula

If A and B are any events whose probabilities are not 0 or 1, then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}$$

The above formula was developed by Thomas Bayes in a book published in 1763 (Barnard, 1958). We will illustrate the application of Bayes' Formula by returning to the meat inspection example. We can use Bayes' Formula to compute $P(\text{E}|\text{Pos})$ for the meat inspection example. To make this calculation, we need to know the *rate* of *E. coli* in the type of meat being inspected. For this example, suppose that *E. coli* is present in 4.5% of all meat samples; that is, *E. coli* has prevalence $P(\text{E}) = .045$. We can then compute $P(\text{E}|\text{Pos})$ as follows:

$$P(\text{E}|\text{Pos}) = \frac{P(\text{Pos}|\text{E})P(\text{E})}{P(\text{Pos}|\text{E})P(\text{E}) + P(\text{Pos}|\text{NE})P(\text{NE})}$$

$$= \frac{(.95)(.045)}{(.95)(.045) + (.01)(1 - .045)} = .817$$

Thus, *E. coli* is truly present in 81.7% of the tested samples in which a positive test result occurs. Also, we can conclude that 18.3% of the tested samples indicated *E. coli* was present when in fact there was no *E. coli* in the meat sample.

EXAMPLE 4.3

A book club classifies members as heavy, medium, or light purchasers, and separate mailings are prepared for each of these groups. Overall, 20% of the members are heavy purchasers, 30% medium, and 50% light. A member is not classified into a group until 18 months after joining the club, but a test is made of the feasibility of using the first 3 months' purchases to classify members. The following percentages are obtained from existing records of individuals classified as heavy, medium, or light purchasers (Table 4.4):

TABLE 4.4
Book club membership classifications

First 3 Months' Purchases	Group (%)		
	Heavy	Medium	Light
0	5	15	60
1	10	30	20
2	30	40	15
3+	55	15	5

If a member purchases no books in the first 3 months, what is the probability that the member is a light purchaser? (*Note:* This table contains "conditional" percentages for each column.)

Solution Using the conditional probabilities in the table, the underlying purchase probabilities, and Bayes' Formula, we can compute this conditional probability.

$$
\begin{aligned}
&P(\text{light}|0) \\
&= \frac{P(0|\text{light})P(\text{light})}{P(0|\text{light})P(\text{light}) + P(0|\text{medium})P(\text{medium}) + P(0|\text{heavy})P(\text{heavy})} \\
&= \frac{(.60)(.50)}{(.60)(.50) + (.15)(.30) + (.05)(.20)} \\
&= .845 \ \blacksquare
\end{aligned}
$$

states of nature
prior probabilities
observable events

likelihoods
posterior probabilities

These examples indicate the basic idea of Bayes' Formula. There is some number k of possible, mutually exclusive, underlying events $A_1,\ldots, A_k$, which are sometimes called the **states of nature.** Unconditional probabilities $P(A_1),\ldots, P(A_k)$, often called **prior probabilities,** are specified. There are m possible, mutually exclusive, **observable events** $B_1,\ldots, B_m$. The conditional probabilities of each observable event given each state of nature, $P(B_i|A_i)$, are also specified, and these probabilities are called **likelihoods.** The problem is to find the **posterior probabilities** $P(A_i|B_i)$. *Prior* and *posterior* refer to probabilities before and after observing an event B_i.

Bayes' Formula

If $A_1,\ldots, A_k$ are mutually exclusive states of nature, and if $B_1,\ldots, B_m$ are m possible, mutually exclusive, observable events, then

$$
\begin{aligned}
P(A_i|B_j) &= \frac{P(B_j|A_i)P(A_i)}{P(B_j|A_1)P(A_1) + P(B_j|A_2)P(A_2) + \cdots + P(B_j|A_k)P(A_k)} \\
&= \frac{P(B_j|A_i)P(A_i)}{\sum_i P(B_j|A_i)P(A_i)}
\end{aligned}
$$

EXAMPLE 4.4

In the manufacture of circuit boards, there are three major types of defective boards. The types of defects, along with the percentage of all circuit boards having these defects, are (1) improper electrode coverage (D_1), 2.8%; (2) plating separation (D_2), 1.2%; and (3) etching problems (D_3), 3.2%. A circuit board will contain at most one of the three defects. Defects can be detected with certainty using destructive testing of the finished circuit boards; however, this is not a very practical method for inspecting a large percentage of the circuit boards. A nondestructive inspection procedure has been developed that has the following outcomes: A_1, which indicates the board has only defect D_1; A_2, which indicates the board has only defect D_2; A_3, which indicates the board has only defect D_3; and A_4, which indicates the board has no defects. The respective likelihoods for the four outcomes of the nondestructive test determined by evaluating a large number of boards known to have exactly one of the three types of defects are given in Table 4.5.

TABLE 4.5
Circuit board defect data

Test Outcome	Type of Defect			
	D_1	D_2	D_3	None
A_1	.90	.06	.02	.02
A_2	.05	.80	.06	.01
A_3	.03	.05	.82	.02
A_4 (no defects)	.02	.09	.10	.95

If a circuit board is tested using the nondestructive test and the outcome indicates no defects (A_4), what are the probabilities that the board has no defect or a D_1, D_2, or D_3 type of defect?

Let D_4 represent the situation in which the circuit board has no defects.

$$P(D_1|A_4) = \frac{P(A_4|D_1)P(D_1)}{P(A_4|D_1)P(D_1) + P(A_4|D_2)P(D_2) + P(A_4|D_3)P(D_3) + P(A_4|D_4)P(D_4)}$$
$$= \frac{(.02)(.028)}{(.02)(.028) + (.09)(.012) + (.10)(.032) + (.95)(.928)} = \frac{.00056}{.88644} = .00063$$

$$P(D_2|A_4) = \frac{P(A_4|D_2)P(D_2)}{P(A_4|D_1)P(D_1) + P(A_4|D_2)P(D_2) + P(A_4|D_3)P(D_3) + P(A_4|D_4)P(D_4)}$$
$$= \frac{(.09)(.012)}{(.02)(.028) + (.09)(.012) + (.10)(.032) + (.95)(.928)} = \frac{.00108}{.88644} = .00122$$

$$P(D_3|A_4) = \frac{P(A_4|D_3)P(D_3)}{P(A_4|D_1)P(D_1) + P(A_4|D_2)P(D_2) + P(A_4|D_3)P(D_3) + P(A_4|D_4)P(D_4)}$$
$$= \frac{(.10)(.032)}{(.02)(.028) + (.09)(.012) + (.10)(.032) + (.95)(.928)} = \frac{.0032}{.88644} = .0036$$

$$P(D_4|A_4) = \frac{P(A_4|D_4)P(D_4)}{P(A_4|D_1)P(D_1) + P(A_4|D_2)P(D_2) + P(A_4|D_3)P(D_3) + P(A_4|D_4)P(D_4)}$$
$$= \frac{(.95)(.928)}{(.02)(.028) + (.09)(.012) + (.10)(.032) + (.95)(.928)} = \frac{.8816}{.88644} = .9945$$

Thus, if the new test indicates that none of the three types of defects is present in the circuit board, there is a very high probability, .9945, that the circuit board in fact is free of defects. In Exercise 4.31, we will ask you to assess the sensitivity of the test for determining the three types of defects. ■

4.6 Variables: Discrete and Continuous

The basic language of probability developed in this chapter deals with many different kinds of events. We are interested in calculating the probabilities associated with both quantitative and qualitative events. For example, we developed techniques that could be used to determine the probability that a machinist selected at random from the workers in a large automotive plant would suffer an accident during an 8-hour shift. These same techniques are also applicable to finding the probability that a machinist selected at random would work more than 80 hours without suffering an accident.

These qualitative and quantitative events can be classified as events (or outcomes) associated with qualitative and quantitative variables. For example, in the automotive plant accident study, the randomly selected machinist's accident report would consist of checking one of the following: No Accident, Minor Accident, or

Major Accident. Thus, the data on 100 machinists in the study would be observations on a qualitative variable because the possible responses are the different categories of accident and are not different in any measurable, numerical amount. Because we cannot predict with certainty what type of accident a particular machinist will suffer, the variable is classified as a **qualitative random variable.** Other examples of qualitative random variables that are commonly measured are political party affiliation, socioeconomic status, the species of insect discovered on an apple leaf, and the brand preferences of customers. There are a finite (and typically quite small) number of possible outcomes associated with any qualitative variable. Using the methods of this chapter, it is possible to calculate the probabilities associated with these events.

qualitative random variable

Many times the events of interest in an experiment are quantitative outcomes associated with a **quantitative random variable,** since the possible responses vary in numerical magnitude. For example, in the automotive plant accident study, the number of consecutive 8-hour shifts between accidents for a randomly selected machinist is an observation on a quantitative random variable. Events of interest, such as the number of 8-hour shifts between accidents for a randomly selected machinist, are observations on a quantitative random variable. Other examples of quantitative random variables are the change in earnings per share of a stock over the next quarter, the length of time a patient is in remission after a cancer treatment, the yield per acre of a new variety of wheat, and the number of persons voting for the incumbent in an upcoming election. The methods of this chapter can be applied to calculate the probability associated with any particular event.

quantitative random variable

There are major advantages to dealing with quantitative random variables. The numerical yardstick underlying a quantitative variable makes the mean and standard deviation (for instance) sensible. With qualitative random variables, the methods of this chapter can be used to calculate the probabilities of various events, and that's about all. With quantitative random variables, we can do much more: We can average the resulting quantities, find standard deviations, and assess probable errors, among other things. Hereafter, we use the term **random variable** to mean quantitative random variable.

random variable

Most events of interest result in numerical observations or measurements. If a quantitative variable measured (or observed) in an experiment is denoted by the symbol y, we are interested in the values that y can assume. These values are called *numerical outcomes.* The number of different plant species per acre in a coal strip mine after a reclamation project is a numerical outcome. The percentage of registered voters who cast ballots in a given election is also a numerical outcome. The quantitative variable y is called a *random variable* because the value that y assumes in a given experiment is a chance or random outcome.

DEFINITION 4.10

When observations on a quantitative random variable can assume only a countable number of values, the variable is called a **discrete random variable.**

Examples of discrete variables are these:

1. Number of bushels of apples per tree of a genetically altered apple variety
2. Change in the number of accidents per month at an intersection after a new signaling device has been installed
3. Number of "dead persons" voting in the last mayoral election in a major midwestern city

Note that it is possible to count the number of values that each of these random variables can assume.

DEFINITION 4.11

When observations on a quantitative random variable can assume any one of the uncountable number of values in a line interval, the variable is called a **continuous random variable.**

For example, the daily maximum temperature in Rochester, New York, can assume any of the infinitely many values on a line interval. It can be 89.6, 89.799, or 89.7611114. Typical continuous random variables are temperature, pressure, height, weight, and distance.

discrete and continuous variables

The distinction between **discrete** and **continuous random variables** is pertinent when we are seeking the probabilities associated with specific values of a random variable. The need for the distinction will be apparent when probability distributions are discussed in later sections of this chapter.

4.7 Probability Distributions for Discrete Random Variables

probability distribution

As previously stated, we need to know the probability of observing a particular sample outcome in order to make an inference about the population from which the sample was drawn. To do this, we need to know the probability associated with each value of the variable y. Viewed as relative frequencies, these probabilities generate a distribution of theoretical relative frequencies called the **probability distribution** of y. Probability distributions differ for discrete and continuous random variables. For discrete random variables, we will compute the probability of specific individual values occurring. For continuous random variables, the probability of an interval of values is the event of interest.

The *probability distribution for a discrete random variable* displays the probability $P(y)$ associated with each value of y. This display can be presented as a table, a graph, or a formula. To illustrate, consider the tossing of two coins in Section 4.2, and let y be the number of heads observed. Then y can take the values 0, 1, or 2. From the data of Table 4.1, we can determine the approximate probability for each value of y, as given in Table 4.6. We point out that the relative frequencies in the table are very close to the theoretical relative frequencies (probabilities), which can be shown to be .25, .50, and .25 using the classical interpretation of probability. If we had employed 2,000,000 tosses of the coins instead of 500, the relative frequencies for $y = 0, 1$, and 2 would be indistinguishable from the theoretical probabilities.

The probability distribution for y, the number of heads in the toss of two coins, is shown in Table 4.7 and is presented graphically in Figure 4.2.

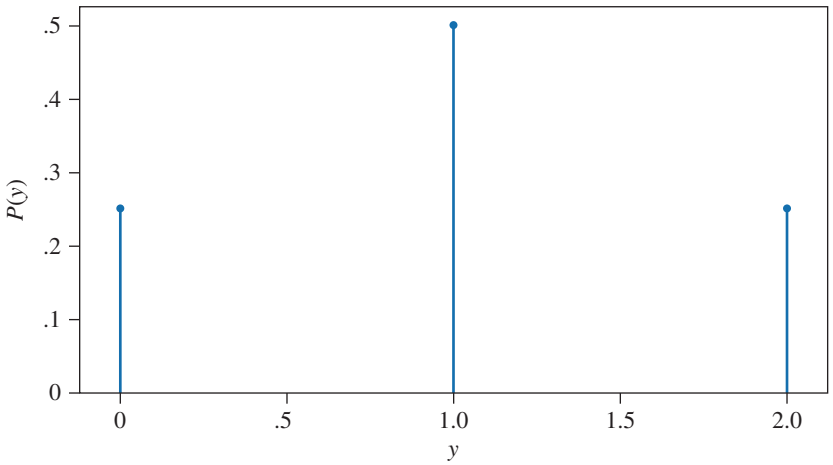
TABLE 4.6
Empirical sampling results for y: the number of heads in 500 tosses of two coins

y	Frequency	Relative Frequency
0	129	.258
1	242	.484
2	129	.258

TABLE 4.7
Probability distribution for the number of heads when two coins are tossed

y	$P(y)$
0	.25
1	.50
2	.25

FIGURE 4.2
Probability distribution for the number of heads when two coins are tossed



The probability distribution for this simple discrete random variable illustrates three important properties of discrete random variables.

Properties of Discrete Random Variables

1. The probability associated with every value of y lies between 0 and 1.
2. The sum of the probabilities for all values of y is equal to 1.
3. The probabilities for a discrete random variable are additive. Hence, the probability that $y = 1$ or 2 is equal to $P(1) + P(2)$.

The relevance of the probability distribution to statistical inference will be emphasized when we discuss the probability distribution for the binomial random variable.

4.8 Two Discrete Random Variables: The Binomial and the Poisson

Many populations of interest to business persons and scientists can be viewed as large sets of 0s and 1s. For example, consider the set of responses of all adults in the United States to the question "Do you favor the development of nuclear energy?" If we disallow "no opinion," the responses will constitute a set of "yes" responses and "no" responses. If we assign a 1 to each yes and a 0 to each no, the population will consist of a set of 0s and 1s, and the sum of the 1s will equal the total number of persons favoring the development. The sum of the 1s divided by the number of adults in the United States will equal the proportion of people who favor the development.

Gallup and Harris polls are examples of the sampling of 0, 1 populations. People are surveyed, and their opinions are recorded. Based on the sample responses, Gallup and Harris estimate the proportions of people in the population who favor some particular issue or possess some particular characteristic.

Similar surveys are conducted in the biological sciences, engineering, and business, but they may be called experiments rather than polls. For example, experiments are conducted to determine the effect of new drugs on small animals, such as rats or mice, before progressing to larger animals and, eventually, to human participants. Many of these experiments bear a marked resemblance to a poll in that

the experimenter records only whether the drug was effective. Thus, if 300 rats are injected with a drug and 230 show a favorable response, the experimenter has conducted a "poll"—a poll of rat reaction to the drug, 230 "in favor" and 70 "opposed."

Similar "polls" are conducted by most manufacturers to determine the fraction of a product that is of good quality. Samples of industrial products are collected before shipment, and each item in the sample is judged "defective" or "acceptable" according to criteria established by the company's quality control department. Based on the number of defectives in the sample, the company can decide whether the product is suitable for shipment. Note that this example, as well as those preceding, has the practical objective of making an inference about a population based on information contained in a sample.

The public opinion poll, the consumer preference poll, the drug-testing experiment, and the industrial sampling for defectives are all examples of a common, frequently conducted sampling situation known as a *binomial experiment.* The binomial experiment is conducted in all areas of science and business and differs from one situation to another only in the nature of objects being sampled (people, rats, electric lightbulbs, oranges). Thus, it is useful to define its characteristics. We can then apply our knowledge of this one kind of experiment to a variety of sampling experiments.

For all practical purposes, the binomial experiment is identical to the coin-tossing example of previous sections. Here n different coins are tossed (or a single coin is tossed n times), and we are interested in the number of heads observed. We assume that the probability of tossing a head on a single trial is π (π may equal .50, as it would for a balanced coin, but in many practical situations, π will take some other value between 0 and 1). We also assume that the outcome for any one toss is unaffected by the results of any preceding tosses. These characteristics can be summarized as shown here.

DEFINITION 4.12

A **binomial experiment** is one that has the following properties:

1. The experiment consists of n identical trials.
2. Each trial results in one of two outcomes. We will label one outcome a success and the other a failure.
3. The probability of success on a single trial is equal to π, and π remains the same from trial to trial.*
4. The trials are independent; that is, the outcome of one trial does not influence the outcome of any other trial.
5. The random variable y is the number of successes observed during the n trials.

EXAMPLE 4.5

An article in the March 5, 1998, issue of ***The New England Journal of Medicine (338:633–639)*** discussed a large outbreak of tuberculosis. One person, called the index patient, was diagnosed with tuberculosis in 1995. The 232 co-workers of the index patient were given a tuberculin screening test. The number of co-workers recording a positive reading on the test was the random variable of interest. Did this study satisfy the properties of a binomial experiment?

*Some textbooks and computer programs use the letter p rather than π. We have chosen π to avoid confusion with p-values, discussed in Chapter 5.

Solution To answer the question, we check each of the five characteristics of the binomial experiment to determine whether they were satisfied.

1. Were there n identical trials? Yes. There were $n = 232$ workers who had approximately equal contact with the index patient.
2. Did each trial result in one of two outcomes? Yes. Each co-worker recorded either a positive or a negative reading on the test.
3. Was the probability of success the same from trial to trial? Yes, if the co-workers had equivalent risk factors and equal exposures to the index patient.
4. Were the trials independent? Yes. The outcome of one screening test was unaffected by the outcomes of the other screening tests.
5. Was the random variable of interest to the experimenter the number of successes y in the 232 screening tests? Yes. The number of co-workers who obtained a positive reading on the screening test was the variable of interest.

All five characteristics were satisfied, so the tuberculin screening test represented a binomial experiment. ■

EXAMPLE 4.6

A large power utility company uses gas turbines to generate electricity. The engineers employed at the company monitor the reliability of each turbine—that is, the probability that the turbine will perform properly under standard operating conditions over a specified period of time. The engineers wanted to estimate the probability a turbine will operate successfully for 30 days after being put into service. The engineers randomly selected 75 of the 100 turbines currently in use and examined the maintenance records. They recorded the number of turbines that did not need repairs during the 30-day time period. Is this a binomial experiment?

Solution Check this experiment against the five characteristics of a binomial experiment.

1. Are there identical trials? The 75 trials could be assumed identical only if the 100 turbines are the same type of turbine, are the same age, and are operated under the same conditions.
2. Does each trial result in one of two outcomes? Yes. Each turbine either does or does not need repairs in the 30-day time period.
3. Is the probability of success the same from trial to trial? No. If we let success denote a turbine "did not need repairs," then the probability of success can change considerably from trial to trial. For example, suppose that 15 of the 100 turbines needed repairs during the 30-day inspection period. Then π, the probability of success for the first turbine examined, would be $85/100 = .85$. If the first trial is a failure (turbine needed repairs), the probability that the second turbine examined did not need repairs is $85/99 = .859$. Suppose that after 60 turbines have been examined, 50 did not need repairs and 10 needed repairs. The probability of success of the next (61st) turbine would be $35/40 = .875$.
4. Were the trials independent? Yes, provided that the failure of one turbine does not affect the performance of any other turbine.

However, the trials may be dependent in certain situations. For example, suppose that a major storm occurs that results in several turbines being damaged. Then the common event, a storm, may result in a common result, the simultaneous failure of several turbines.

5. Was the random variable of interest to the engineers the number of successes in the 75 trials? Yes. The number of turbines not needing repairs during the 30-day period was the random variable of interest.

This example shows how the probability of success can change substantially from trial to trial in situations in which the sample size is a relatively large portion of the total population size. This experiment does not satisfy the properties of a binomial experiment. ■

Note that very few real-life situations satisfy perfectly the requirements stated in Definition 4.12, but for many, the lack of agreement is so small that the binomial experiment still provides a very good model for reality.

Having defined the binomial experiment and suggested several practical applications, we now examine the probability distribution for the binomial random variable y, the number of successes observed in n trials. Although it is possible to approximate $P(y)$, the probability associated with a value of y in a binomial experiment, by using a relative frequency approach, it is easier to use a general formula for binomial probabilities.

Formula for Computing $P(y)$ in a Binomial Experiment

The probability of observing y successes in n trials of a binomial experiment is

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}$$

where

n = number of trials
π = probability of success on a single trial
$1-\pi$ = probability of failure on a single trial
y = number of successes in n trials
$n! = n(n-1)(n-2)\cdots(3)(2)(1)$

As indicated in the box, the notation $n!$ (referred to as n factorial) is used for the product

$$n! = n(n-1)(n-2)\ldots(3)(2)(1)$$

For $n = 3$,

$$n! = 3! = (3)(3-1)(3-2) = (3)(2)(1) = 6$$

Similarly, for $n = 4$,

$$4! = (4)(3)(2)(1) = 24$$

We also note that 0! is defined to be equal to 1.

To see how the formula for binomial probabilities can be used to calculate the probability for a specific value of y, consider the following examples.

EXAMPLE 4.7

A new variety of turf grass has been developed for use on golf courses, with the goal of obtaining a germination rate of 85%. To evaluate the grass, 20 seeds are planted in a greenhouse so that each seed will be exposed to identical conditions. If the 85% germination rate is correct, what is the probability that 18 or more of the 20 seeds will germinate?

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}$$

and substituting for $n = 20$, $\pi = .85$, $y = 18$, 19, and 20, we obtain

$$P(y = 18) = \frac{20!}{18!(20-18)!}(.85)^{18}(1-.85)^{20-18} = 190(.85)^{18}(.15)^2 = .229$$

$$P(y = 19) = \frac{20!}{19!(20-19)!}(.85)^{19}(1-.85)^{20-19} = 20(.85)^{19}(.15)^1 = .137$$

$$P(y = 20) = \frac{20!}{20!(20-20)!}(.85)^{20}(1-.85)^{20-20} = (.85)^{20} = .0388$$

$$P(y \geq 18) = P(y = 18) + P(y = 19) + P(y = 20) = .405 \blacksquare$$

The calculations in Example 4.7 entail a considerable amount of effort even though n was only 20. For those situations involving a large value of n, we can use computer software to make the exact calculations. The following commands in R will compute the binomial probabilities:

1. To calculate $P(X = 18)$, use the command **dbinom(18, 20, .85)**
2. To calculate $P(X \leq 17)$, use the command **pbinom (17, 20, .85)**
3. To calculate $P(X \geq 18)$, use the command **1 − pbinom(17, 20, .85)**

Later in this chapter, the normal approximation to the binomial will be discussed. This approximation yields fairly accurate results and does not require the use of a computer.

EXAMPLE 4.8

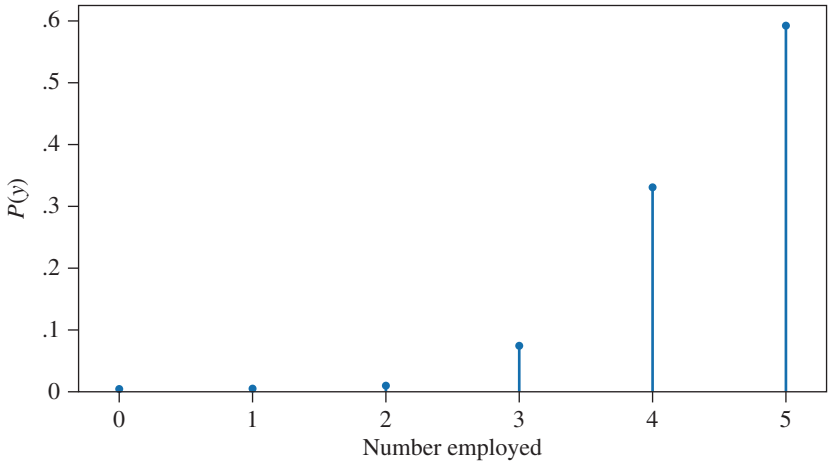
Suppose that a sample of households is randomly selected from all the households in the city in order to estimate the percentage in which the head of the household is unemployed. To illustrate the computation of a binomial probability, suppose that the unknown percentage is actually 10% and that a sample of $n = 5$ (we select a small sample to make the calculation manageable) is selected from the population. What is the probability that all five heads of households are employed?

Solution We must carefully define which outcome we wish to call a success. For this example, we define a success as being employed. Then the probability of success when one person is selected from the population is $\pi = .9$ (because the proportion unemployed is .1). We wish to find the probability that $y = 5$ (all five are employed) in five trials.

$$\begin{aligned} P(y = 5) &= \frac{5!}{5!(5-5)!}(.9)^5(1-.9)^{5-5} \\ &= \frac{5!}{5!0!}(.9)^5(.1)^0 \\ &= \frac{(5)(4)(3)(2)(1)}{(5)(4)(3)(2)(1)(1)}(.9)^5(.1)^0 \\ &= (.9)^5 = .590 \end{aligned}$$

The binomial probability distribution for $n = 5, \pi = .9$ is shown in Figure 4.3. The probability of observing five employed in a sample of five is shown to be 0.59 in Figure 4.3.

FIGURE 4.3
The binomial probability distribution for $n = 5, \pi = .9$



EXAMPLE 4.9

Refer to Example 4.8 and calculate the probability that exactly one person in the sample of five households is unemployed. What is the probability of one or fewer being unemployed?

Solution Since y is the number of employed in the sample of five, one unemployed person would correspond to four employed ($y = 4$). Then

$$\begin{aligned} P(4) &= \frac{5!}{4!(5-4)!}(.9)^4(.1)^1 \\ &= \frac{(5)(4)(3)(2)(1)}{(4)(3)(2)(1)(1)}(.9)^4(.1) \\ &= 5(.9)^4(.1) \\ &= .328 \end{aligned}$$

Thus, the probability of selecting four employed heads of households in a sample of five is .328, or roughly one chance in three.

The outcome "one or fewer unemployed" is the same as the outcome "4 or 5 employed." Since y represents the number employed, we seek the probability that $y = 4$ or 5. Because the values associated with a random variable represent mutually exclusive events, the probabilities for discrete random variables are additive. Thus, we have

$$\begin{aligned} P(y = 4 \text{ or } 5) &= P(4) + P(5) \\ &= .328 + .590 \\ &= .918 \end{aligned}$$

Thus, the probability that a random sample of five households will yield either four or five employed heads of households is .918. This high probability is consistent with our intuition: We would expect the number of employed in the sample to be large if 90% of all heads of households in the city are employed. ■

Like any relative frequency histogram, a binomial probability distribution possesses a mean, μ, and a standard deviation, σ. Although we omit the derivations, we give the formulas for these parameters.

Mean and Standard Deviation of the Binomial Probability Distribution

$$\mu = n\pi \quad \text{and} \quad \sigma = \sqrt{n\pi(1-\pi)}$$

where π is the probability of success in a given trial and n is the number of trials in the binomial experiment.

If we know π and the sample size, n, we can calculate μ and σ to locate the average value and describe the variability for a particular binomial probability distribution. Thus, we can quickly determine those values of y that are probable and those that are improbable.

EXAMPLE 4.10

We will consider the turf grass seed example to illustrate the calculation of the mean and standard deviation. Suppose the company producing the turf grass takes a sample of 20 seeds on a regular basis to monitor the quality of the seeds. If the germination rate of the seeds stays constant at 85%, then the average number of seeds that will germinate in the sample of 20 seeds is

$$\mu = n\pi = 20(.85) = 17$$

with a standard deviation of

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{20(.85)(1-.85)} = 1.60$$

Suppose we examine the germination records of a large number of samples of 20 seeds each. If the germination rate has remained constant at 85%, then the average number of seeds that germinate should be close to 17 per sample. If in a particular sample of 20 seeds we determine that only 12 had germinated, would the germination rate of 85% seem consistent with our results? Using a computer software program, we can generate the probability distribution for the number of seeds that germinate in the sample of 20 seeds, as shown in Figures 4.4(a) and 4.4(b).

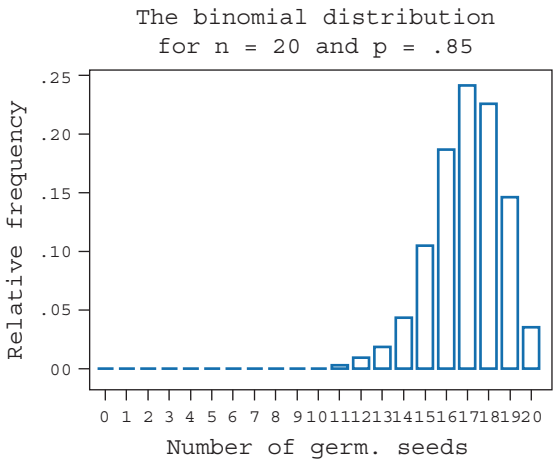
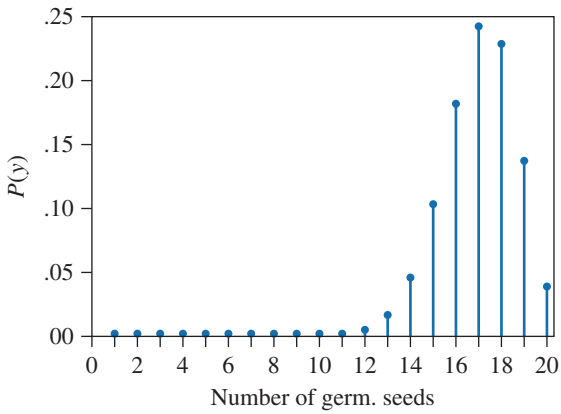


FIGURE 4.4(a)
The binomial distribution for $n = 20$ and $p = .85$

FIGURE 4.4(b)
The binomial distribution for $n = 20$ and $p = .85$



A software program was used to generate Figure 4.4(a). Many such packages place rectangles centered at each of the possible integer values of the binomial random variable, as shown in Figure 4.4(a), even though there is zero probability for any value but the integers to occur. This results in a distorted representation of the binomial distribution. A more appropriate display of the distribution is given in Figure 4.4(b).

Although the distribution is tending toward left skewness (see Figure 4.4(b)), the Empirical Rule should work well for this relatively mound-shaped distribution. Thus, $y = 12$ seeds is more than three standard deviations less than the mean number of seeds, $\mu = 17$; it is highly improbable that in 20 seeds we would obtain only 12 germinated seeds if π really is equal to .85. The germination rate is most likely a value considerably less than .85. ■

EXAMPLE 4.11

A cable TV company is investigating the feasibility of offering a new service in a large midwestern city. In order for the proposed new service to be economically viable, it is necessary that at least 50% of its current subscribers add the new service. A survey of 1,218 customers reveals that 516 would add the new service. Do you think the company should expend the capital to offer the new service in this city?

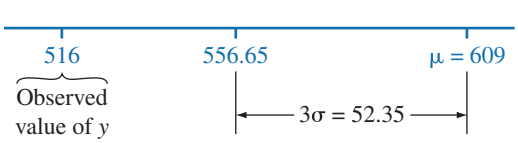
Solution In order to be economically viable, the company needs at least 50% of its current customers to subscribe to the new service. Is $y = 516$ out of 1,218 too small a value of y to imply a value of π (the proportion of current customers who would add new service) equal to .50 or larger? If $\pi = .5$,

$$\mu = n\pi = 1{,}218(.5) = 609$$
$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{1{,}218(.5)(1-.5)} = 17.45$$

and $3\sigma = 52.35$.

You can see from Figure 4.5 that $y = 516$ is more than 3σ, or 52.35, less than $\mu = 609$, the value of μ if π really equalled .5. Thus, the observed number of customers in the sample who would add the new service is much too small if the

FIGURE 4.5
Location of the observed value of y ($y = 516$) relative to μ



number of current customers who would not add the service in fact is 50% or more of all customers. Consequently, the company concluded that offering the new service was not a good idea. ■

The purpose of this section is to present the binomial probability distribution so you can see how binomial probabilities are calculated and so you can calculate them for small values of n, if you wish. In practice, n is usually large (in national surveys, sample sizes as large as 1,500 are common), and the computation of the binomial probabilities is tedious. Later in this chapter, we will present a simple procedure for obtaining approximate values of the probabilities we need in making inferences. In order to obtain very accurate calculations when n is large, we recommend using a computer software program. (See Section 4.16.)

Poisson distribution

In 1837, S. D. Poisson developed a discrete probability distribution, suitably called the **Poisson distribution,** which has as one of its important applications the modeling of events of a particular time over a unit of time or space—for example, the number of automobiles arriving at a toll booth during a given 5-minute period of time. The event of interest would be an arriving automobile, and the unit of time would be 5 minutes. A second example would be the situation in which an environmentalist measures the number of PCB particles discovered in a liter of water sampled from a stream contaminated by an electronics production plant. The event would be a PCB particle is discovered. The unit of space would be 1 liter of sampled water.

Let y be the number of events occurring during a fixed time interval of length t or a fixed region R of area or volume $m(R)$. Then the probability distribution of y is Poisson, provided certain conditions are satisfied:

1. Events occur one at a time; two or more events do not occur precisely at the same time or in the same space.
2. The occurrence of an event in a given period of time or region of space is independent of the occurrence of the event in a nonoverlapping time period or region of space; that is, the occurrence (or nonoccurrence) of an event during one period or in one region does not affect the probability of an event occurring at some other time or in some other region.
3. The expected number of events during one period or in one region, μ, is the same as the expected number of events in any other period or region.

Although these assumptions seem somewhat restrictive, many situations appear to satisfy these conditions. For example, the number of arrivals of customers at a checkout counter, parking lot toll booth, inspection station, or garage repair shop during a specified time interval can often be modeled by a Poisson distribution. Similarly, the number of clumps of algae of a particular species observed in a unit volume of lake water could be approximated by a Poisson probability distribution.

Assuming that the above conditions hold, the Poisson probability of observing y events in a unit of time or space is given by the formula

$$P(y) = \frac{\mu^y e^{-\mu}}{y!}$$

where e is a naturally occurring constant approximately equal to 2.71828 (in fact, $e = 2 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots$), $y! = y(y-1)(y-2)\cdots(1)$, and μ is the average value of y. Table 14 in the Appendix gives Poisson probabilities for various values of the parameter μ.

EXAMPLE 4.12

A large industrial plant is being planned in a rural area. As a part of the environmental impact statement, a team of wildlife scientists is surveying the number and types of small mammals in the region. Let y denote the number of field mice captured in a trap over a 24-hour period. Suppose that y has a Poisson distribution with $\mu = 2.3$; that is, the average number of field mice captured per trap is 2.3. What is the probability of finding exactly four field mice in a randomly selected trap? What is the probability of finding at most four field mice in a randomly selected trap? What is the probability of finding more than four field mice in a randomly selected trap?

Solution The probability that a trap contains exactly four field mice is computed to be

$$P(y = 4) = \frac{e^{-2.3}(2.3)^4}{4!} = \frac{(.1002588)(27.9841)}{24} = .1169$$

Alternatively, we could use Table 14 in the Appendix. We read from the table with $\mu = 2.3$ and $y = 4$ that $P(y = 4) = .1169$.

The probability of finding at most four field mice in a randomly selected trap is, using the values from Table 14, with $\mu = 2.3$

$$\begin{aligned} P(y \le 4) &= P(y = 0) + P(y = 1) + P(y = 2) + P(y = 3) + P(y = 4) \\ &= .1003 + .2306 + .2652 + .2033 + .1169 = .9163. \end{aligned}$$

The probability of finding more than four field mice in a randomly selected trap, using the idea of complementary events, is

$$P(y > 4) = 1 - P(y \le 4) = 1 - .9163 = .0837$$

Thus, it is a very unlikely event to find five or more field mice in a trap.

The Poisson probabilities can be computed using the following R commands.

$P(y = 4) =$ **dpois(4, 2.3)** $= .1169022$
$P(y \le 3) =$ **ppois(3, 2.3)** $= .7993471$
$P(y > 4) = 1 - P(y \le 4) = 1 -$ **ppois(4, 2.3)** $= .08375072$

When n is large and π is small in a binomial experiment, $n \ge 100$, $\pi \le .01$, and $n\pi \le 20$, the Poisson distribution provides an reasonable approximation to the binomial distribution. In applying the Poisson approximation to the binomial distribution, use $\mu = n\pi$. ■

EXAMPLE 4.13

In observing patients administered a new drug product in a properly conducted clinical trial, the number of persons experiencing a particular side effect might be quite small. Suppose π (the probability a person experiences a side effect to the drug) is .001 and 1,000 patients in the clinical trial received the drug. Compute the probability that none of a random sample of $n = 1{,}000$ patients administered the drug experiences a particular side effect (such as damage to a heart valve) when $\pi = .001$.

Solution The number of patients, y, experiencing the side effect would have a binomial distribution with $n = 1{,}000$ and $\pi = .001$. The mean of the binomial distribution is $\mu = n\pi = 1{,}000(.001) = 1$. Applying the Poisson probability distribution with $\mu = 1$, we have

$$P(y = 0) = \frac{(1)^0 e^{-1}}{0!} = e^{-1} = \frac{1}{2.71828} = .367879$$

(Note also from Table 14 in the Appendix that the entry corresponding to $y = 0$ and $\mu = 1$ is .3679.) ■

For the calculation in Example 4.13, it is easy to compute the exact binomial probability and then compare the results to the Poisson approximation. With $n = 1{,}000$ and $\pi = .001$, we obtain the following.

$$P(y = 0) = \frac{1{,}000!}{0!(1{,}000 - 0)!}(.001)^0(1 - .001)^{1{,}000} = (.999)^{1{,}000} = .367695$$

The Poisson approximation was accurate to the third decimal place.

EXAMPLE 4.14

Suppose that after a clinical trial of a new medication involving 1,000 patients, no patient experienced a side effect to the drug. Would it be reasonable to infer that less than .1% of the entire population would experience this side effect while taking the drug?

Solution Certainly not. We computed the probability of observing $y = 0$ in $n = 1{,}000$ trials, assuming $\pi = .001$ (i.e., assuming .1% of the population would experience the side effect), to be .368. Because this probability is quite large, it would not be wise to infer that $\pi < .001$. Rather, we would conclude that there is not sufficient evidence to contradict the assumption that π is .001 or larger. ■

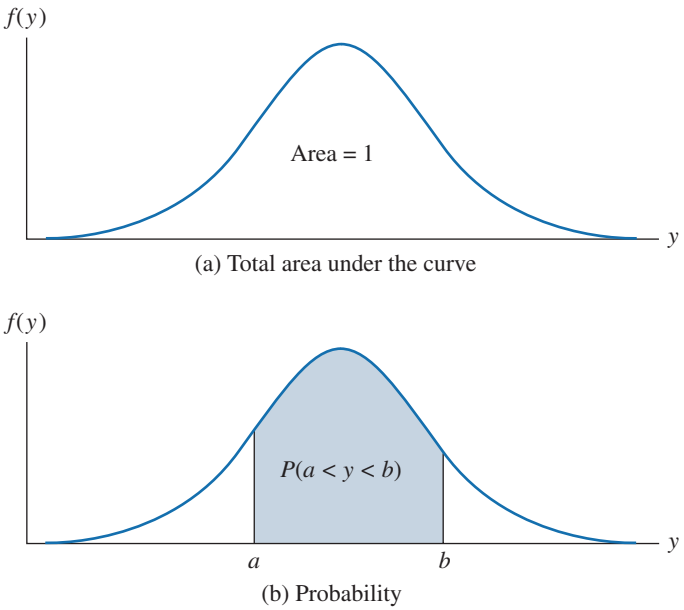
4.9 Probability Distributions for Continuous Random Variables

Discrete random variables (such as the binomial) have possible values that are distinct and separate, such as 0 or 1 or 2 or 3. Other random variables are most usefully considered to be *continuous:* Their possible values form a whole interval (or range, or continuum). For instance, the 1-year return per dollar invested in a common stock could range from 0 to some quite large value. In practice, virtually all random variables assume a discrete set of values; the return per dollar of a million-dollar common-stock investment could be \$1.06219423 or \$1.06219424 or \$1.06219425 or.... However, when there are many possible values for a random variable, it is sometimes mathematically useful to treat the random variable as continuous.

Theoretically, then, a continuous random variable is one that can assume values associated with infinitely many points in a line interval. We state, without elaboration, that it is impossible to assign a small amount of probability to each value of y (as was done for a discrete random variable) and retain the property that the probabilities sum to 1.

To overcome this difficulty, we revert to the concept of the relative frequency histogram of Chapter 3, where we talked about the probability of y falling in a given interval. Recall that the relative frequency histogram for a population containing a large number of measurements will almost be a smooth curve because the number of class intervals can be made large and the width of the intervals

FIGURE 4.6
Probability distribution for a continuous random variable



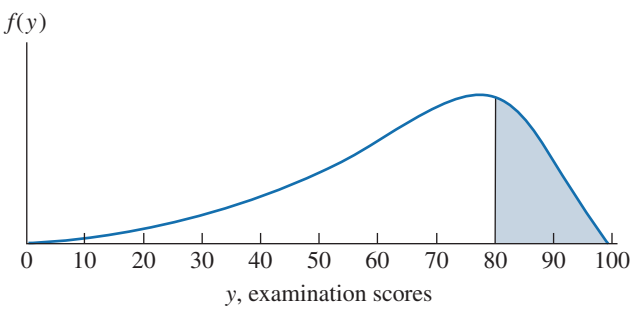
can be decreased to a very small value. Thus, we envision a smooth curve that provides a model for the population relative frequency distribution generated by repeated observation of a continuous random variable. This will be similar to the curve shown in Figure 4.6.

Recall that the histogram relative frequencies are proportional to areas over the class intervals and that these areas possess a probabilistic interpretation. Thus, if a measurement is randomly selected from the set, the probability that it will fall in an interval is proportional to the histogram area above the interval. Since a population is the whole (100%, or 1), we want the total area under the probability curve to equal 1. If we let the total area under the curve equal 1, then areas over intervals are exactly equal to the corresponding probabilities.

The graph for the probability distribution for a continuous random variable is shown in Figure 4.7. The ordinate (height of the curve) for a given value of y is denoted by the symbol $f(y)$. Many people are tempted to say that $f(y)$, like $P(y)$ for the binomial random variable, designates the probability associated with the continuous random variable y. However, as we mentioned before, it is impossible to assign a probability to each of the infinitely many possible values of a continuous random variable. Thus, all we can say is that $f(y)$ represents the height of the probability distribution for a given value of y.

The probability that a continuous random variable falls in an interval—say, between two points a and b—follows directly from the probabilistic interpretation

FIGURE 4.7
Hypothetical probability distribution for student examination scores

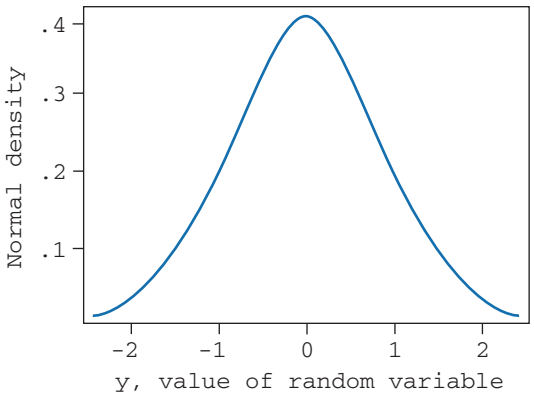


given to the area over an interval for the relative frequency histogram (Section 3.3) and is equal to the area under the curve over the interval a to b, as shown in Figure 4.6. This probability is written $P(a < y < b)$.

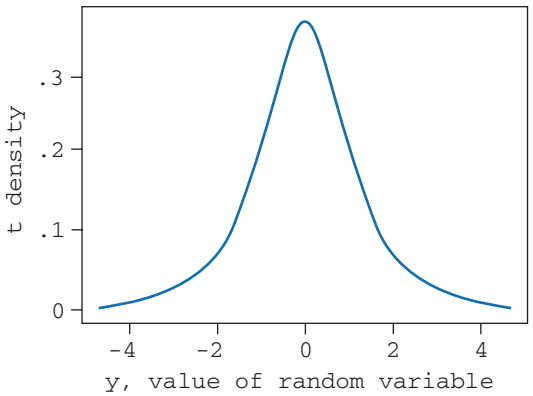
There are curves of many shapes that can be used to represent the population relative frequency distribution for measurements associated with a continuous random variable. Fortunately, the areas for many of these curves have been tabulated and are ready for use. Thus, if we know that student examination scores possess a particular probability distribution, as in Figure 4.7, and if areas under the curve have been tabulated, we can find the probability that a particular student will score more than 80 by looking up the tabulated area, which is shaded in Figure 4.7.

Figure 4.8 depicts four important probability distributions that will be used extensively in the following chapters. Which probability distribution we use in a particular situation is very important because probability statements are determined by the area under the curve. As can be seen in Figure 4.8, we would obtain very different answers depending on which distribution is selected. For example, the probability the random variable takes on a value less than 5.0 is essentially 1.0 for the probability distributions in Figures 4.8(a) and (b) but is .584 and .947 for the probability distributions in Figures 4.8(c) and (d), respectively. In some situations, we will not know exactly the distribution for

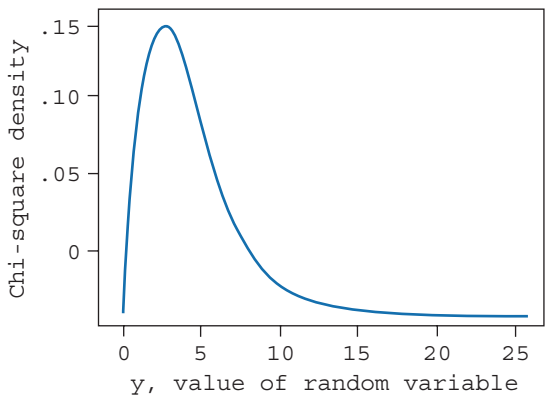
FIGURE 4.8 Probability distributions of normal, t, chi-square, and F



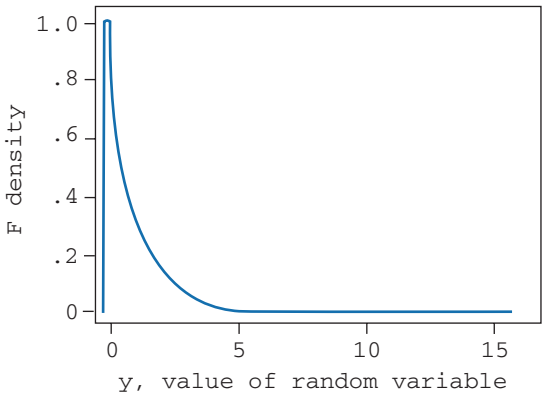
(a) Density of the standard normal distribution



(b) Density of the t(df = 3) distribution



(c) Density of the chi-square (df = 5) distribution



(d) Density of the F(df = 2, 6) distribution

the random variable in a particular study. In these situations, we can use the observed values for the random variable to construct a relative frequency histogram, which is a sample estimate of the true probability frequency distribution. As far as statistical inferences are concerned, the selection of the *exact* shape of the probability distribution for a continuous random variable is not crucial in many cases because most of our inference procedures are insensitive to the exact specification of the shape.

We will find that data collected on continuous variables often possess a nearly bell-shaped frequency distribution, such as depicted in Figure 4.8(a). A continuous variable (the normal) and its probability distribution (bell-shaped curve) provide a good model for these types of data. The normally distributed variable is also very important in statistical inference. We will study the normal distribution in detail in the next section.

4.10 A Continuous Probability Distribution: The Normal Distribution

normal curve

Many variables of interest, including several statistics to be discussed in later sections and chapters, have mound-shaped frequency distributions that can be approximated by using a **normal curve.** For example, the distribution of total scores on the Brief Psychiatric Rating Scale for outpatients having a current history of repeated aggressive acts is mound-shaped. Other practical examples of mound-shaped distributions are social perceptiveness scores of preschool children selected from a particular socioeconomic background, psychomotor retardation scores for patients with circular-type manic-depressive illness, milk yields for cattle of a particular breed, and perceived anxiety scores for residents of a community. Each of these mound-shaped distributions can be approximated with a normal curve.

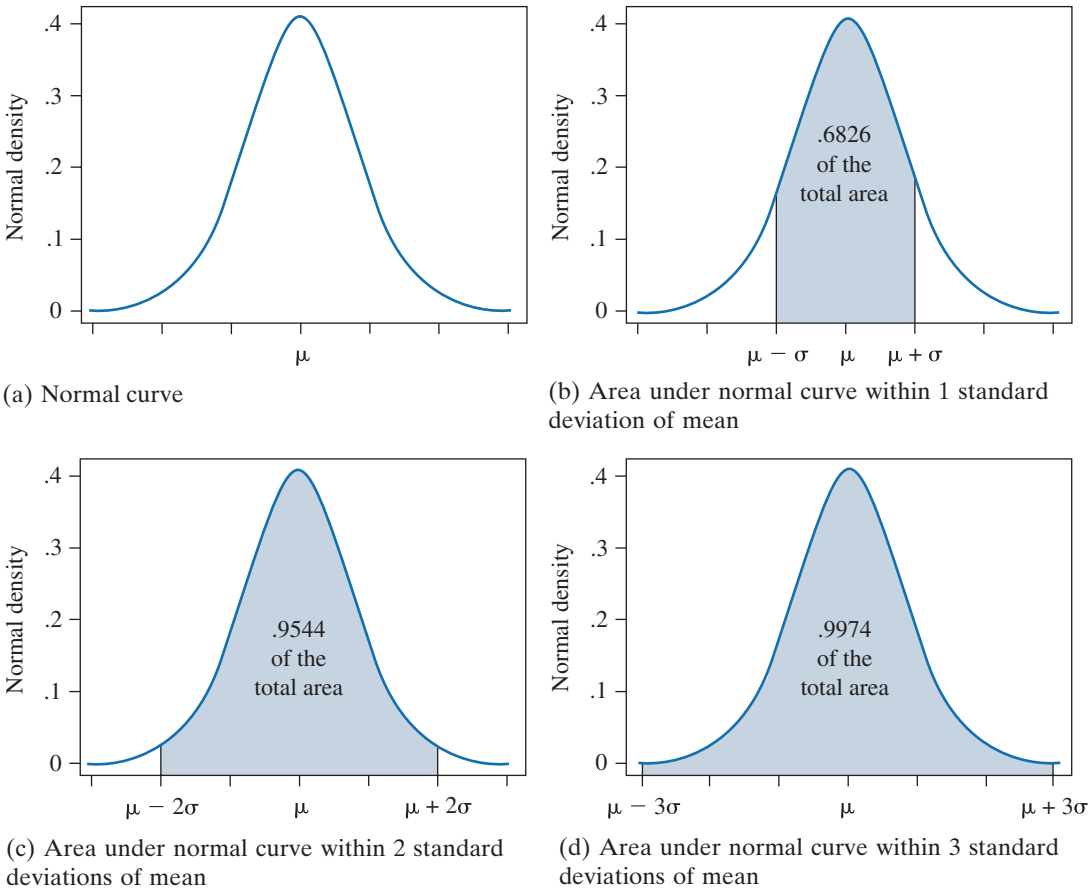
Since the normal distribution has been well tabulated, areas under a normal curve—which correspond to probabilities—can be used to approximate probabilities associated with the variables of interest in our experimentation. Thus, the normal random variable and its associated distribution play an important role in statistical inference.

The relative frequency histogram for the normal random variable, called the *normal curve* or *normal probability distribution,* is a smooth, bell-shaped curve. Figure 4.9(a) shows a normal curve. If we let y represent the normal random variable, then the height of the probability distribution for a specific value of y is represented by $f(y)$.* The probabilities associated with a normal curve form the basis for the Empirical Rule.

As we see from Figure 4.9(a), the normal probability distribution is bell-shaped and symmetrical about the mean μ. Although the normal random variable y may theoretically assume values from $-\infty$ to $+\infty$, we know from the Empirical Rule that approximately all the measurements are within 3 standard deviations (3σ) of μ. From the Empirical Rule, we also know that if we select a measurement at random from a population of measurements that possesses a mound-shaped distribution, the probability is approximately .68 that the measurement will lie within 1 standard

*For the normal distribution, $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2}$, where μ and σ are the mean and standard deviation, respectively, of the population of y-values.

FIGURE 4.9 Normal distribution



(a) Normal curve

(b) Area under normal curve within 1 standard deviation of mean

(c) Area under normal curve within 2 standard deviations of mean

(d) Area under normal curve within 3 standard deviations of mean

deviation of its mean (see Figure 4.9(b)). Similarly, we know that the probability is approximately .954 that a value will lie in the interval $\mu \pm 2\sigma$ and .997 in the interval $\mu \pm 3\sigma$ (see Figures 4.9(c) and (d)). What we do not know, however, is the probability that the measurement will be within 1.65 standard deviations of its mean, or within 2.58 standard deviations of its mean. The procedure we are going to discuss in this section will enable us to calculate the probability that a measurement falls within any distance of the mean μ for a normal curve.

Because there are many different normal curves (depending on the parameters μ and σ), it might seem to be an impossible task to tabulate areas (probabilities) for all normal curves, especially if each curve requires a separate table. Fortunately, this is not the case. By specifying the probability that a variable y lies within a certain number of standard deviations of its mean (just as we did in using the Empirical Rule), we need only one table of probabilities.

area under a normal curve

Table 1 in the Appendix gives the **area under a normal curve** to the left of a value y that is z standard deviations ($z\sigma$) away from the mean (see Figure 4.10). The area shown by the shading in Figure 4.10 is the probability listed in Table 1 in the Appendix. Values of z to the nearest tenth are listed along the left-hand column of the table, with z to the nearest hundredth along the top of the table. To find the probability that a normal random variable will lie to the left of a point 1.65 standard deviations above the mean, we look up the table entry corresponding to $z = 1.65$. This probability is .9505 (see Figure 4.11).

FIGURE 4.10
Area under a normal curve as given in Appendix Table 1

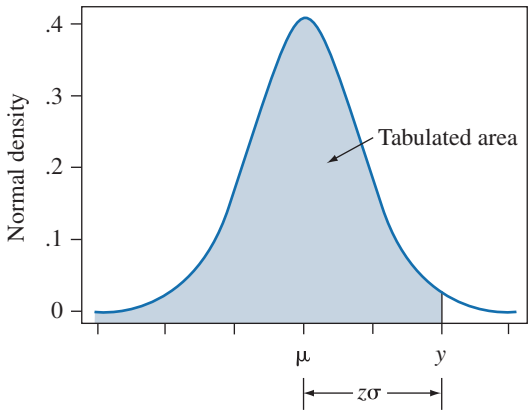
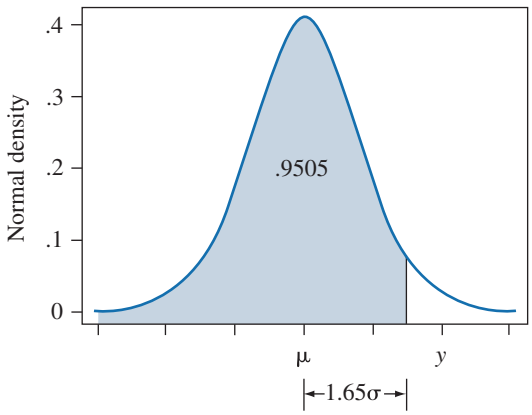


FIGURE 4.11
Area under a normal curve from μ to a point 1.65 standard deviations above the mean



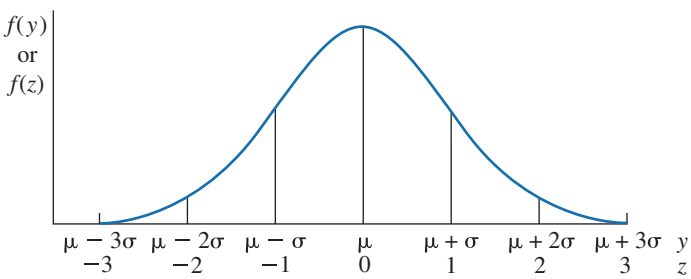
To determine the probability that a measurement will be less than some value y, we first calculate the number of standard deviations that y lies away from the mean by using the formula

$$z = \frac{y - \mu}{\sigma}$$

z-score

The value of z computed using this formula is sometimes referred to as the ***z*-score** associated with the y-value. Using the computed value of z, we determine the appropriate probability by using Table 1 in the Appendix. Note that we are merely coding the value y by subtracting μ and dividing by σ. (In other words, $y = z\sigma + \mu$.) Figure 4.12 illustrates the values of z corresponding to specific values of y. Thus, a value of y that is 2 standard deviations below (to the left of) μ corresponds to $z = -2$.

FIGURE 4.12
Relationship between specific values of y and $z = (y - \mu)/\sigma$

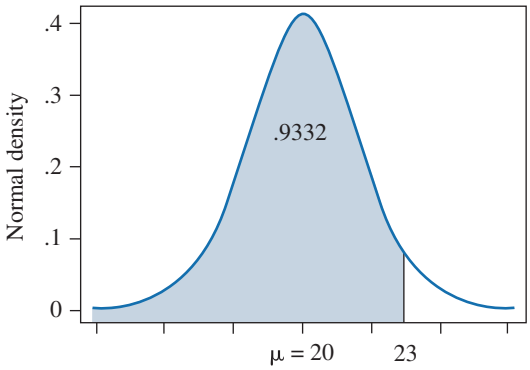


EXAMPLE 4.15

Consider a normal distribution with $\mu = 20$ and $\sigma = 2$. Determine the probability that a measurement will be less than 23.

Solution When first working problems of this type, it might be a good idea to draw a picture so that you can see the area in question, as we have in Figure 4.13.

FIGURE 4.13
Area less than $y = 23$ under normal curve, with $\mu = 20, \sigma = 2$



To determine the area under the curve to the left of the value $y = 23$, we first calculate the number of standard deviations $y = 23$ lies away from the mean.

$$z = \frac{y - \mu}{s} = \frac{23 - 20}{2} = 1.5$$

Thus, $y = 23$ lies 1.5 standard deviations above $\mu = 20$. Referring to Table 1 in the Appendix, we find the area corresponding to $z = 1.5$ to be .9332. This is the probability that a measurement is less than 23. ■

EXAMPLE 4.16

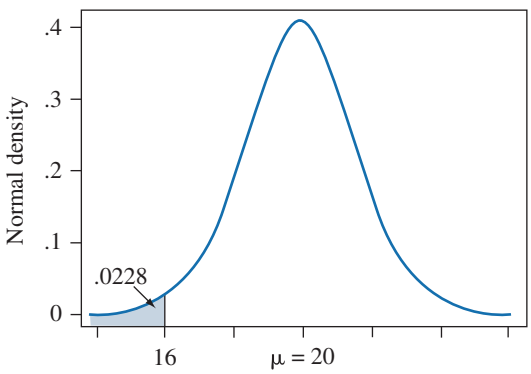
For the normal distribution of Example 4.15 with $\mu = 20$ and $\sigma = 2$, find the probability that y will be less than 16.

Solution In determining the area to the left of 16, we use

$$z = \frac{y - \mu}{\sigma} = \frac{16 - 20}{2} = -2$$

We find the appropriate area from Table 1 to be .0228; thus, .0228 is the probability that a measurement is less than 16. The area is shown in Figure 4.14.

FIGURE 4.14
Area less than $y = 16$ under normal curve, with $\mu = 20, \sigma = 2$



■

EXAMPLE 4.17

A high accumulation of ozone gas in the lower atmosphere at ground level is air pollution and can be harmful to people, animals, crops, and various materials. Elevated levels above the national standard may cause lung and respiratory disorders. Nitrogen oxides and hydrocarbons are known as the chief "precursors" of ozone. These compounds react in the presence of sunlight to produce ozone. The sources of these precursor pollutants include cars, trucks, power plants, and factories. Large industrial areas and cities with heavy summer traffic are the main contributors to ozone formation. The United States Environmental Protection Agency (EPA) has developed procedures for measuring vehicle emission levels of nitrogen oxide. Let P denote the amount of this pollutant in a randomly selected automobile in Houston, Texas. Suppose the distribution of P can be adequately modeled by a normal distribution with a mean level of $\mu = 70$ ppb (parts per billion) and a standard deviation of $\sigma = 13$ ppb.

a. What is the probability that a randomly selected vehicle will have an emission level less than 60 ppb?
b. What is the probability that a randomly selected vehicle will have an emission level greater than 90 ppb?
c. What is the probability that a randomly selected vehicle will have an emission level between 60 and 90 ppb?

Solution We begin by drawing pictures of the areas that we are looking for (Figures 4.15(a)–(c)). To answer part (a), we must compute the z-value corresponding to the y-value of 60. The value $y = 60$ corresponds to a z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{60 - 70}{13} = -.77$$

From Table 1, the area to the left of 60 is .2206 (see Figure 4.15(a)). Alternatively, we could use the R command **pnorm (− .77)**.

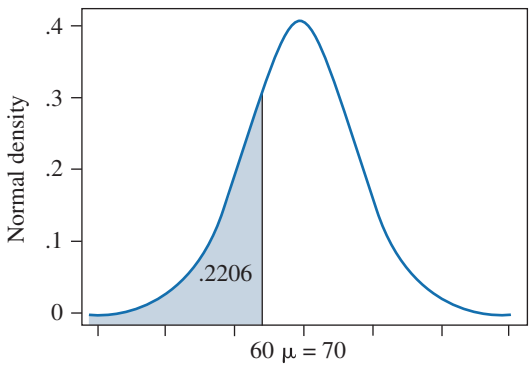


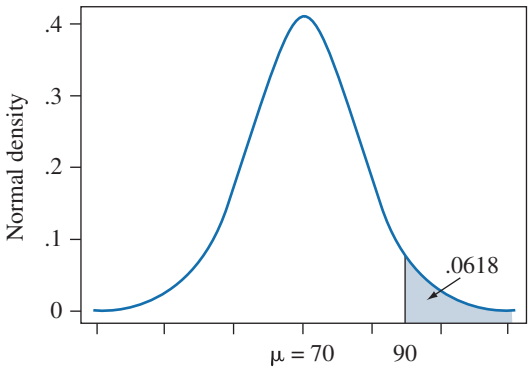
FIGURE 4.15(a) Area less than $y = 60$ under normal curve, with $\mu = 70, \sigma = 13$

To answer part (b), the value $y = 90$ corresponds to a z-score of

$$z = \frac{y - \mu}{s} = \frac{90 - 70}{13} = 1.54$$

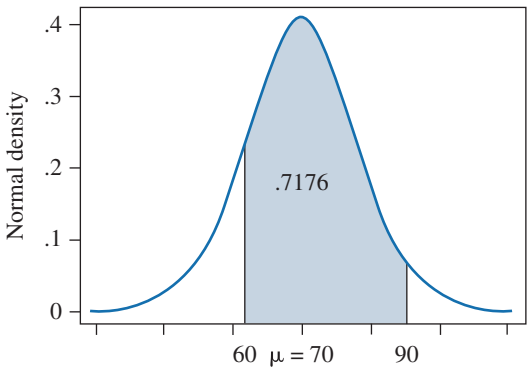
so from Table 1 we obtain .9382, the tabulated area less than 90. Thus, the area greater than 90 must be $1 - .9382 = .0618$, since the total area under the curve is 1 (see Figure 4.15(b)). Alternatively, $1 -$ **pnorm(1.54)** $= .0618$.

FIGURE 4.15(b)
Area greater than $y = 90$ under normal curve, with $\mu = 70, \sigma = 13$



To answer part (c), we can use our results from (a) and (b). The area between two values y_1 and y_2 is determined by finding the difference between the areas to the left of the two values, (see Figure 4.15(c)). We found that the area less than 60 is .2206 and the area less than 90 is .9382. Hence, the area between 60 and 90 is .9382 − .2206 = .7176. We can thus conclude that 22.06% of inspected vehicles will have nitrogen oxide levels less than 60 ppb, 6.18% of inspected vehicles will have nitrogen oxide levels greater than 90 ppb, and 71.76% of inspected vehicles will have nitrogen oxide levels between 60 ppb and 90 ppb.

FIGURE 4.15(c)
Area between 60 and 90 under normal curve, with $\mu = 70, \sigma = 13$



■

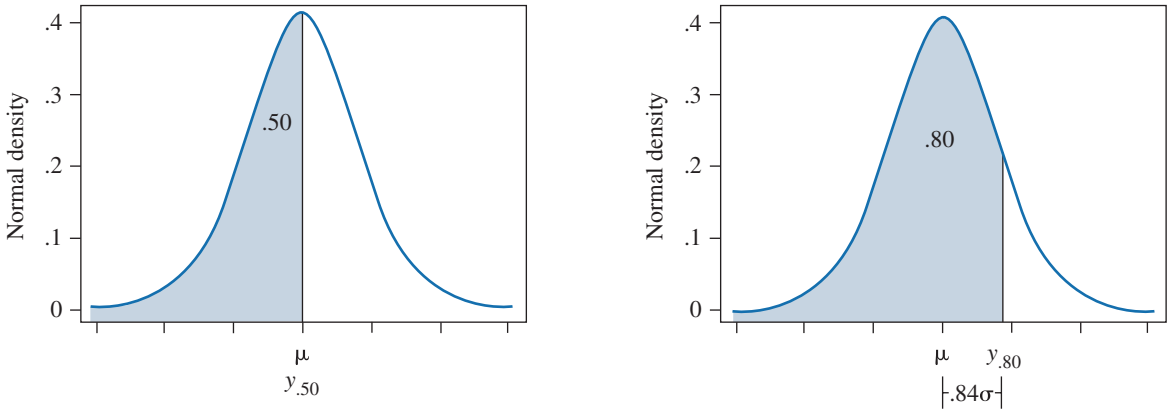
100*p*th percentile

An important aspect of the normal distribution is that we can easily find the percentiles of the distribution. The **100*p*th percentile** of a distribution is that value, y_p, such that $100p\%$ of the population values fall below y_p and $100(1 - p)\%$ are above y_p. For example, the median of a population is the 50th percentile, $y_{.50}$, and the quartiles are the 25th and 75th percentiles. The normal distribution is symmetric, so the median and the mean are the same value, $y_{.50} = \mu$ (see Figure 4.16(a)).

To find the percentiles of the standard normal distribution, we reverse our use of Table 1. To find the 100*p*th percentile, z_p, we find the probability p in Table 1 and then read out its corresponding number, z_p, along the margins of the table. For example, to find the 80th percentile, $z_{.80}$, we locate the probability $p = .8000$ in Table 1. The value nearest to .8000 is .7995, which corresponds to a z-value of 0.84. Thus, $z_{.80} = 0.84$ (see Figure 4.16(b)). Now, to find the 100*p*th percentile, y_p, of a normal distribution with mean μ and standard deviation σ, we need to apply the reverse of our standardization formula,

$$y_p = \mu + z_p\sigma$$

FIGURE 4.16 Mean, median, 80th percentile of normal distribution



(a) For the normal curve, the mean and median agree

(b) The 80th percentile for the normal curve

Suppose we wanted to determine the 80th percentile of a population having a normal distribution with $\mu = 55$ and $\sigma = 3$. We have determined that $z_{.80} = 0.84$; thus, the 80th percentile for the population would be $y_{.80} = 55 + (.84)(3) = 57.52$. Alternatively, we could use the *R* command **qnorm(.8, 55, 3)**.

EXAMPLE 4.18

A State of Texas environmental agency, using the vehicle inspection process described in Example 4.17, is going to offer a reduced vehicle license fee to those vehicles having very low emission levels. As a preliminary pilot project, it will offer this incentive to the group of vehicle owners having the best 10% of emission levels. What emission level should the agency use in order to identify the best 10% of all emission levels?

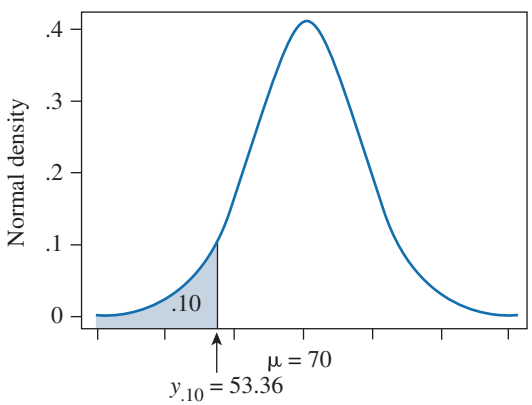
Solution The best 10% of all emission levels would be the 10% having the lowest emission levels, as depicted in Figure 4.17.

To find the 10th percentile (see Figure 4.17), we first find $z_{.10}$ in Table 1. Since .1003 is the value nearest .1000 and its corresponding z-value is -1.28, we take $z_{.10} = -1.28$. We then compute

$$y_{.10} = \mu + z_{.10}\sigma = 70 + (-1.28)(13) = 70 - 16.64 = 53.36$$

Thus, 10% of the vehicles have emissions less than 53.36 ppb. Alternatively, $y_{.10}$ = **qnorm(.1, 70, 13)**.

FIGURE 4.17
The 10th percentile for a normal curve, with $\mu = 70, \sigma = 13$



■

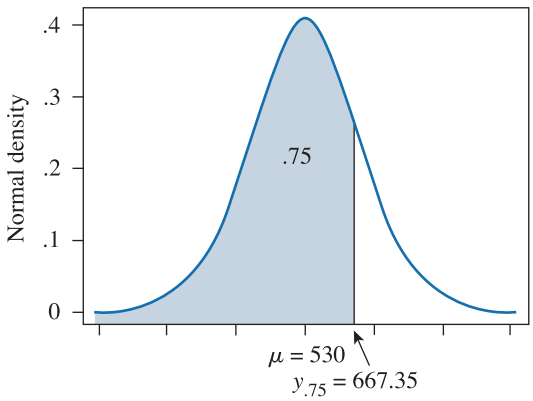
EXAMPLE 4.19

An analysis of income tax returns from the previous year indicates that for a given income classification, the amount of money owed to the government over and above the amount paid in the estimated tax vouchers for the first three payments is approximately normally distributed with a mean of $530 and a standard deviation of $205. Find the 75th percentile for this distribution of measurements. The government wants to target that group of returns having the largest 25% of amounts owed.

Solution We need to determine the 75th percentile, $y_{.75}$ (Figure 4.18). From Table 1, we find $z_{.75} = .67$ because the probability nearest .7500 is .7486, which corresponds to a z-score of .67. We then compute

$$y_{.75} = \mu + z_{.75}\sigma = 530 + (.67)(205) = 667.35$$

FIGURE 4.18
The 75th percentile for a normal curve, with $\mu = 530, \sigma = 205$



Thus, 25% of the tax returns in this classification exceed $667.35 in the amount owed the government. ■

4.11 Random Sampling

Thus far in the text, we have discussed random samples and introduced various sampling schemes in Chapter 2. What is the importance of random sampling? We must know how the sample was selected so we can determine probabilities associated with various sample outcomes. The probabilities of samples selected *in a random manner* can be determined, and we can use these probabilities to make inferences about the population from which the sample were drawn.

Sample data selected in a nonrandom fashion are frequently distorted by a *selection bias.* A selection bias exists whenever there is a systematic tendency to overrepresent or underrepresent some part of the population. For example, a survey of households conducted during the week entirely between the hours of 9 A.M. and 5 P.M. would be severely biased toward households with at least one member at home. Hence, any inferences made from the sample data would be biased toward the attributes or opinions of those families with at least one member at home and may not be truly representative of the population of households in the region.

random sample

Now we turn to a definition of a **random sample** of n measurements selected from a population containing N measurements $(N > n)$. (*Note:* This is a simple random sample, as discussed in Chapter 2. Since most of the random samples discussed in this text will be simple random samples, we'll drop the adjective unless needed for clarification.)

DEFINITION 4.13

A sample of n measurements selected from a population is said to be a **random sample** if every different sample of size n from the population has an equal probability of being selected.

EXAMPLE 4.20

A study of crimes related to handguns is being planned for the 10 largest cities in the United States. The study will randomly select 2 of the 10 largest cities for an in-depth study following the preliminary findings. The population of interest is the 10 largest cities (C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}). List all possible different samples consisting of 2 cities that could be selected from the population of 10 cities. Give the probability associated with each sample in a random sample of $n = 2$ cities selected from the population.

Solution All possible samples are listed in Table 4.8.

TABLE 4.8
Samples of size 2

Sample	Cities	Sample	Cities	Sample	Cities
1	C_1, C_2	16	C_2, C_9	31	C_5, C_6
2	C_1, C_3	17	C_2, C_{10}	32	C_5, C_7
3	C_1, C_4	18	C_3, C_4	33	C_5, C_8
4	C_1, C_5	19	C_3, C_5	34	C_5, C_9
5	C_1, C_6	20	C_3, C_6	35	C_5, C_{10}
6	C_1, C_7	21	C_3, C_7	36	C_6, C_7
7	C_1, C_8	22	C_3, C_8	37	C_6, C_8
8	C_1, C_9	23	C_3, C_9	38	C_6, C_9
9	C_1, C_{10}	24	C_3, C_{10}	39	C_6, C_{10}
10	C_2, C_3	25	C_4, C_5	40	C_7, C_8
11	C_2, C_4	26	C_4, C_6	41	C_7, C_9
12	C_2, C_5	27	C_4, C_7	42	C_7, C_{10}
13	C_2, C_6	28	C_4, C_8	43	C_8, C_9
14	C_2, C_7	29	C_4, C_9	44	C_8, C_{10}
15	C_2, C_8	30	C_4, C_{10}	45	C_9, C_{10}

Now let us suppose that we select a random sample of $n = 2$ cities from the 45 possible samples. The sample selected is called a *random* sample if every sample has an equal probability, 1/45, of being selected. ■

random number table

One of the simplest and most reliable ways to select a random sample of n measurements from a population is to use a table of random numbers (see Table 13 in the Appendix). **Random number tables** are constructed in such a way that, no matter where you start in the table and no matter in which direction you move, the digits occur randomly and with equal probability. Thus, if we wished to choose a random sample of $n = 10$ measurements from a population containing 100 measurements, we could label the measurements in the population from 0 to 99 (or 1 to 100). Then by referring to Table 13 in the Appendix and choosing a random starting point, the next 10 two-digit numbers going across the page would indicate the labels of the particular measurements to be included in the random sample. Similarly, by moving up or down the page, we would also obtain a random sample.

This listing of all possible samples is feasible only when both the sample size n and the population size N are small. We can determine the number, M, of distinct

samples of size n that can be selected from a population of N measurements using the following formula:

$$M = \frac{N!}{n!(N-n)!}$$

In Example 4.20, we had $N = 10$ and $n = 2$. Thus,

$$M = \frac{10!}{2!(10-2)!} = \frac{10!}{2!8!} = 45$$

The value of M becomes very large even when N is fairly small. For example, if $N = 50$ and $n = 5$, then $M = 2{,}118{,}760$. Thus, it would be very impractical to list all 2,118,760 possible samples consisting of $n = 5$ measurements from a population of $N = 50$ measurements and then randomly select one of the samples. In practice, we construct a list of elements in the population by assigning a number from 1 to N to each element in the population, called the *sampling frame*. We then randomly select n integers from the integers $(1, 2, \ldots, N)$ by using a table of random numbers (see Table 13 in the Appendix) or by using a computer program. Most statistical software programs contain routines for randomly selecting n integers from the integers $(1, 2, \ldots, N)$, where $N > n$. For example, the R command **sample (seq(l:N), n, replace = False)** would produce a random sample of n integers from the collection of integers $1, 2, \ldots, N$.

EXAMPLE 4.21

The school board in a large school district has decided to test for illegal drug use among those high school students participating in extracurricular activities. Because these tests are very expensive, they have decided to institute a random testing procedure. Every week 20 students will be randomly selected from the 850 high school students participating in extracurricular activities, and drug tests will be performed. Refer to Table 13 in the Appendix or use a computer software program to determine which students should be tested.

Solution Using the list of all 850 students participating in extracurricular activities, we label the students from 0 to 849 (or, equivalently, from 1 to 850). Then, referring to Table 13 in the Appendix, we select a starting point (close your eyes and pick a point in the table). Suppose we selected line 1, column 3. Going down the page in Table 13, we select the first 20 three-digit numbers between 000 and 849. We would obtain the following 20 numbers:

015	110	482	333
255	564	526	463
225	054	710	337
062	636	518	224
818	533	524	055

These 20 numbers identify the 20 students that are to be included in the first week of drug testing. We would repeat the process in subsequent weeks using a new starting point. The R command **sample(seq(1:850), 20, replace = False)** would produce a random sample of 20 integers from the integers 1 to 850. ■

A telephone directory is often used in selecting people to participate in surveys or pools, especially in surveys related to economics or politics. In the 1936 presidential campaign, Franklin Roosevelt was running as the Democratic candidate against the Republican candidate, Governor Alfred Landon of Kansas. This was

a difficult time for the nation; the country had not yet recovered from the Great Depression of the early 1930s, and there were still 9 million people unemployed.

The *Literary Digest* set out to sample the voting public and predict the winner of the election. Using names and addresses taken from telephone books and club memberships, the *Literary Digest* sent out 10 million questionnaires and got 2.4 million back. Based on the responses to the questionnaire, the *Digest* predicted a Landon victory by 57% to 43%.

At this time, George Gallup was starting his survey business. He conducted two surveys. The first one, based on 3,000 people, predicted what the results of the *Digest* survey would be long before the *Digest* results were published; the second survey, based on 50,000, was used to forecast *correctly* the Roosevelt victory.

How did Gallup correctly predict what the *Literary Digest* survey would predict and then, with another survey, correctly predict the outcome of the election? Where did the *Literary Digest* go wrong? The first problem was a severe selection bias. By taking the names and addresses from telephone directories and club memberships, its survey systematically excluded the poor. Unfortunately for the *Digest,* the vote was split along economic lines; the poor gave Roosevelt a large majority, whereas the rich tended to vote for Landon. A second reason for the error could be due to a *nonresponse bias.* Because only 20% of the 10 million people returned their surveys and approximately half of those responding favored Landon, one might suspect that maybe the nonrespondents had different preferences than did the respondents. This was in fact true.

How then does one achieve a random sample? Careful planning and a certain amount of ingenuity are required to have even a decent chance to approximate random sampling. This is especially true when the universe of interest involves people. People can be difficult to work with; they have a tendency to discard mail questionnaires and refuse to participate in personal interviews. Unless we are very careful, the data we obtain may be full of biases having unknown effects on the inferences we are attempting to make.

We do not have sufficient time to explore the topic of random sampling further in this text; entire courses at the undergraduate and graduate levels can be devoted to sample-survey research methodology. The important point to remember is that data from a random sample will provide the foundation for making statistical inferences in later chapters. Random samples are not easy to obtain, but with care, we can avoid many potential biases that could affect the inferences we make. References providing detailed discussions on how to properly conduct a survey were given in Chapter 2.

4.12 Sampling Distributions

We discussed several different measures of central tendency and variability in Chapter 3 and distinguished between numerical descriptive measures of a population (parameters) and numerical descriptive measures of a sample (statistics). Thus, μ and σ are parameters, whereas $\bar{y}$ and s are statistics.

The numerical value of a sample statistic cannot be predicted exactly in advance. Even if we knew that a population mean μ was \$216.37 and that the population standard deviation σ was \$32.90—even if we knew the complete population distribution—we could not say that the sample mean $\bar{y}$ would be exactly equal to \$216.37. A sample statistic is a random variable; it is subject to random variation because it is based on a random sample of measurements selected from the population of interest. Also, like any other random variable, a sample statistic has a

probability distribution. We call the probability distribution of a sample statistic the *sampling distribution* of that statistic. Stated differently, the sampling distribution of a statistic is the population of all possible values for that statistic.

The actual mathematical derivation of sampling distributions is one of the basic problems of mathematical statistics. We will illustrate how the sampling distribution for $\bar{y}$ can be obtained for a simplified population. Later in the chapter, we will present several general results.

EXAMPLE 4.22

The sample $\bar{y}$ is to be calculated from a random sample of size 2 taken from a population consisting of 10 values (2, 3, 4, 5, 6, 7, 8, 9, 10, 11). Find the sampling distribution of $\bar{y}$, based on a random sample of size 2.

Solution One way to find the sampling distribution is by counting. There are 45 possible samples of 2 items selected from the 10 items. These are shown in Table 4.9.

TABLE 4.9
List of values for the sample mean, $\bar{y}$

Sample	Value of $\bar{y}$	Sample	Value of $\bar{y}$	Sample	Value of $\bar{y}$
2, 3	2.5	3, 10	6.5	6, 7	6.5
2, 4	3	3, 11	7	6, 8	7
2, 5	3.5	4, 5	4.5	6, 9	7.5
2, 6	4	4, 6	5	6, 10	8
2, 7	4.5	4, 7	5.5	6, 11	8.5
2, 8	5	4, 8	6	7, 8	7.5
2, 9	5.5	4, 9	6.5	7, 9	8
2, 10	6	4, 10	7	7, 10	8.5
2, 11	6.5	4, 11	7.5	7, 11	9
3, 4	3.5	5, 6	5.5	8, 9	8.5
3, 5	4	5, 7	6	8, 10	9
3, 6	4.5	5, 8	6.5	8, 11	9.5
3, 7	5	5, 9	7	9, 10	9.5
3, 8	5.5	5, 10	7.5	9, 11	10
3, 9	6	5, 11	8	10, 11	10.5

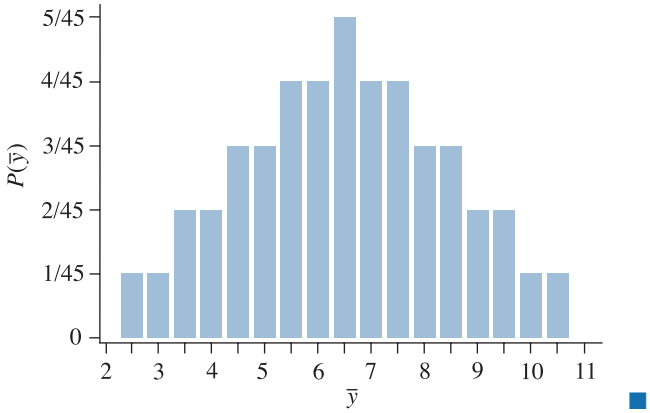
Assuming each sample of size 2 is equally likely, it follows that the sampling distribution for $\bar{y}$ based on $n = 2$ observations selected from the population {2, 3, 4, 5, 6, 7, 8, 9, 10, 11} is as indicated in Table 4.10.

TABLE 4.10
Sampling distribution for $\bar{y}$

$\bar{y}$	$P(\bar{y})$	$\bar{y}$	$P(\bar{y})$
2.5	1/45	7	4/45
3	1/45	7.5	4/45
3.5	2/45	8	3/45
4	2/45	8.5	3/45
4.5	3/45	9	2/45
5	3/45	9.5	2/45
5.5	4/45	10	1/45
6	4/45	10.5	1/45
6.5	5/45		

The sampling distribution is shown as a graph in Figure 4.19. Note that the distribution is symmetric, with a mean of 6.5 and a standard deviation of approximately 2.0 (the range divided by 4).

FIGURE 4.19
Sampling distribution for $\bar{y}$



Example 4.22 illustrates for a very small population that we could in fact enumerate every possible sample of size 2 selected from the population and then compute all possible values of the sample mean. The next example will illustrate the properties of the sample mean, $\bar{y}$, when sampling from a larger population. This example will illustrate that the behavior of $\bar{y}$ as an estimator of μ depends on the sample size, n. Later in this chapter, we will illustrate the effect of the shape of the population distribution on the sampling distribution of $\bar{y}$.

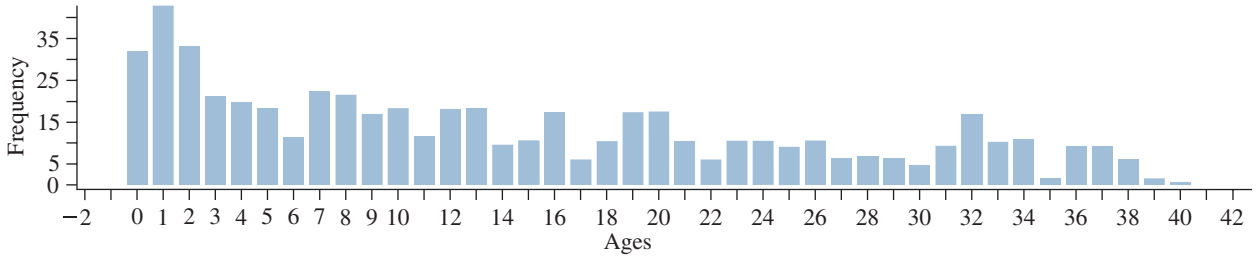
EXAMPLE 4.23

In this example, the population values are known, and, hence, we can compute the exact values of the population mean, μ, and population standard deviation, σ. We will then examine the behavior of $\bar{y}$ based on samples of size $n = 5$, 10, and 25 selected from the population. The population consists of 500 pennies from which we compute the age of each penny: Age = 2015 − Date on penny. The histogram of the 500 ages is displayed in Figure 4.20(a). The shape is skewed to the right with a very long right tail. The mean and standard deviation are computed to be $\mu = 13.468$ years and $\sigma = 11.164$ years. In order to generate the sampling distribution of $\bar{y}$ for $n = 5$, we would need to generate all possible samples of size $n = 5$ and then compute the $\bar{y}$ from each of these samples. This would be an enormous task, since there are 255,244,687,600 possible samples of size 5 that could be selected from a population of 500 elements. The number of possible samples of size 10 or 25 is so large it makes even the national debt look small. Thus, we will use a computer program to select 25,000 samples of size 5 from the population of 500 pennies. For example, the first sample consists of pennies with ages 4, 12, 26, 16, and 9. The sample mean $\bar{y} = (4 + 12 + 26 + 16 + 9)/5 = 13.4$. We repeat 25,000 times the process of selecting 5 pennies; recording their ages, y_1, y_2, y_3, y_4, y_5; and then computing $\bar{y} = (y_1 + y_2 + y_3 + y_4 + y_5)/5$. The 25,000 values for $\bar{y}$ are then plotted in a frequency histogram, called the *sampling distribution* of $\bar{y}$ for $n = 5$. A similar procedure is followed for samples of size $n = 10$ and $n = 25$. The sampling distributions obtained are displayed in Figures 4.20(b)–(d).

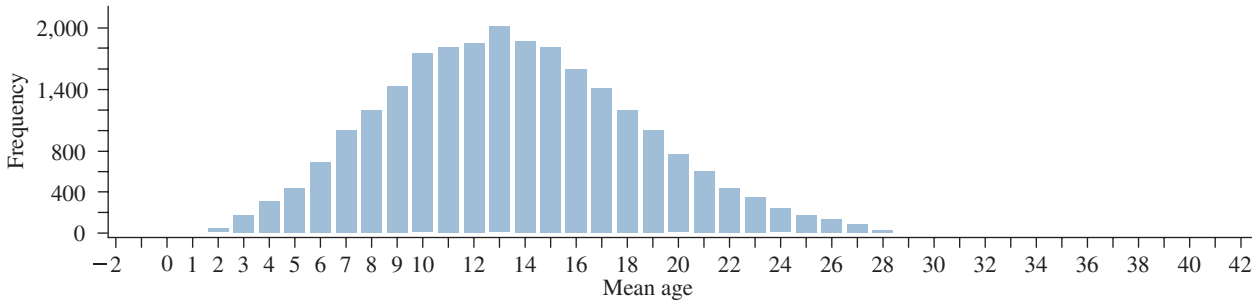
Note that all three sampling distributions have nearly the same central value, approximately 13.5. (See Table 4.11.) The mean values of $\bar{y}$ for the three samples are nearly the same as the population mean, $\mu = 13.468$. In fact, if we had generated all possible samples for all three values of n, the mean of the possible values of $\bar{y}$ would agree exactly with μ.

The next characteristic to notice about the three histograms is their shape. All three are somewhat symmetric in shape, achieving a nearly normal distribution

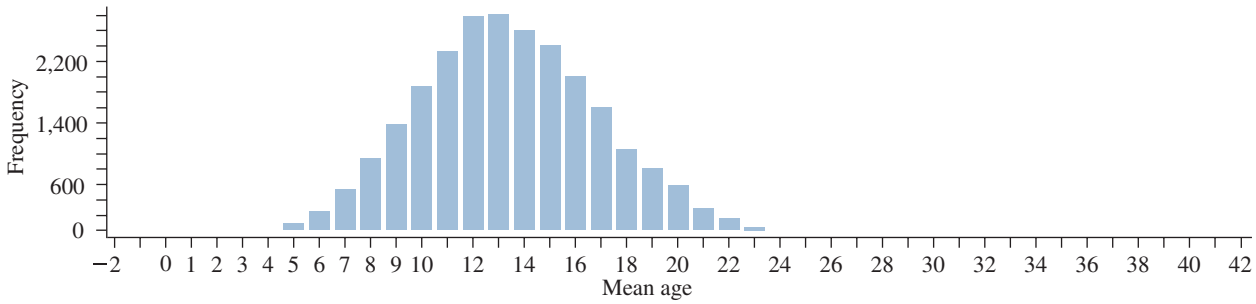
FIGURE 4.20 Sampling distribution of $\bar{y}$ for $n = 1, 5, 10, 25$



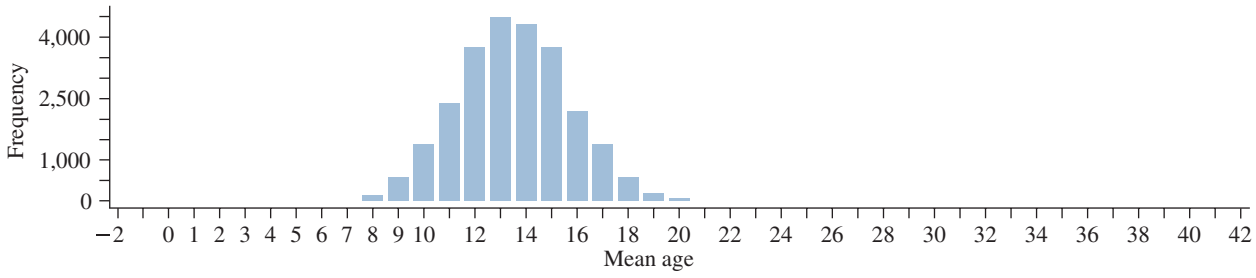
(a) Histogram of ages for 500 pennies



(b) Sampling distribution of $\bar{y}$ for $n = 5$



(c) Sampling distribution of $\bar{y}$ for $n = 10$



(d) Sampling distribution of $\bar{y}$ for $n = 25$

TABLE 4.11
Means and standard deviations for the sampling distributions of $\bar{y}$

Sample Size	Mean of $\bar{y}$	Standard Deviation of $\bar{y}$	$11.1638/\sqrt{n}$
1 (Population)	13.468 (μ)	11.1638 (σ)	11.1638
5	13.485	4.9608	4.9926
10	13.438	3.4926	3.5303
25	13.473	2.1766	2.2328

shape when $n = 25$. However, the histogram for $\bar{y}$ based on samples of size $n = 5$ is more spread out than the histogram based on $n = 10$, which, in turn, is more spread out than the histogram based on $n = 25$. When n is small, we are much more likely to obtain a value of $\bar{y}$ far from μ than when n is large. What causes this increased dispersion in the values of $\bar{y}$? A single extreme y, either large or small relative to μ, in the sample has a greater influence on the size of $\bar{y}$ when n is small than when n is large. Thus, sample means based on small n are less accurate in their estimation of μ than are their large-sample counterparts.

Table 4.11 contains summary statistics for the sampling distribution of $\bar{y}$. The sampling distribution of $\bar{y}$ has mean $\mu_{\bar{y}}$ and standard deviation $\sigma_{\bar{y}}$, which are related to the population mean, μ, and standard deviation, σ, by the following relationships:

$$\mu_{\bar{y}} = \mu \quad \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

From Table 4.11, we note that the three sampling deviations have means that are approximately equal to the population mean. Also, the three sampling deviations have standard deviations that are approximately equal to $\sigma/\sqrt{n}$. If we had generated all possible values of $\bar{y}$, then the standard deviation of $\bar{y}$ would equal $\sigma/\sqrt{n}$ exactly. This quantity, $\sigma_{\bar{y}} = \sigma/\sqrt{n}$, is called the **standard error of $\bar{y}$.** ■

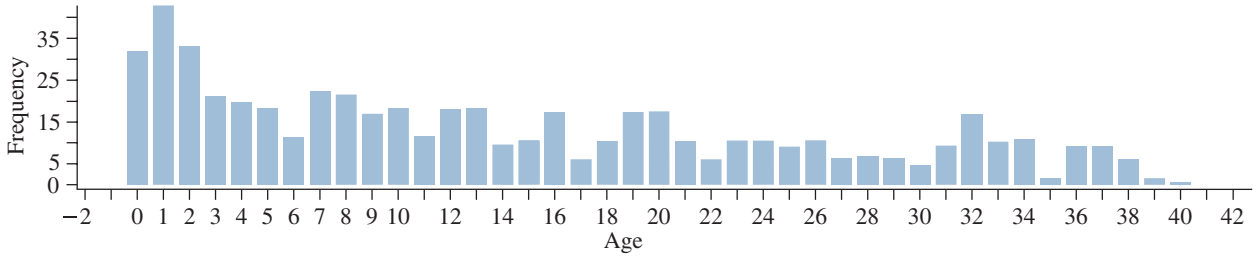
standard error of $\bar{y}$

Quite a few of the more common sample statistics, such as the sample median and the sample standard deviation, have sampling distributions that are nearly normal for moderately sized values of n. We can observe this behavior by computing the sample median and sample standard deviation from each of the three sets of 25,000 samples ($n = 5$, 10, 25) selected from the population of 500 pennies. The resulting sampling distributions are displayed in Figures 4.21(b)–(d), for the sample median, and Figures 4.22(b)–(d), for the sample standard deviation. The sampling distributions of both the median and the standard deviation are more highly skewed in comparison to the sampling distribution of the sample mean. In fact, the value of n at which the sampling distributions of the sample median and standard deviation have a nearly normal shape is much larger than the value required for the sample mean. A series of theorems in mathematical statistics called the **Central Limit Theorems** provide theoretical justification for our approximating the true sampling distribution of many sample statistics with the normal distribution. We will discuss one such theorem for the sample mean. Similar theorems exist for the sample median, sample standard deviation, and sample proportion.

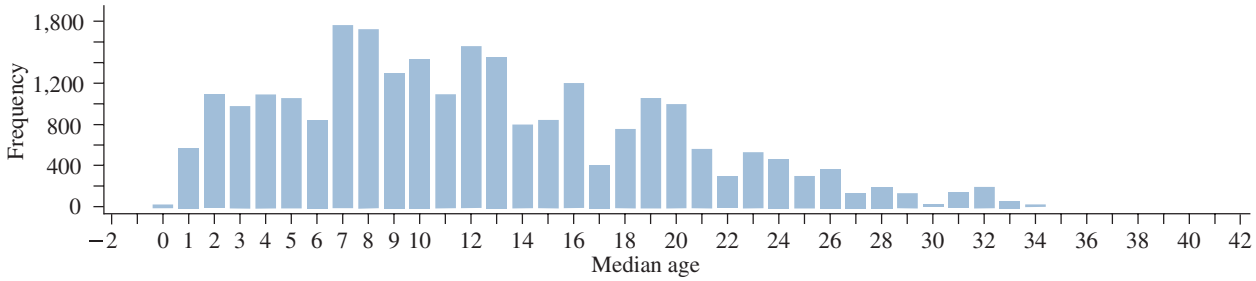
Central Limit Theorems

Figure 4.20 illustrates the Central Limit Theorem. Figure 4.20(a) displays the distribution of the measurements y in the population from which the samples are to be drawn. No specific shape was required for these measurements for the Central Limit Theorem to be validated. Figures 4.20(b)–(d) illustrate the sampling distribution for the sample mean $\bar{y}$ when n is 5, 10, and 25, respectively. We note that

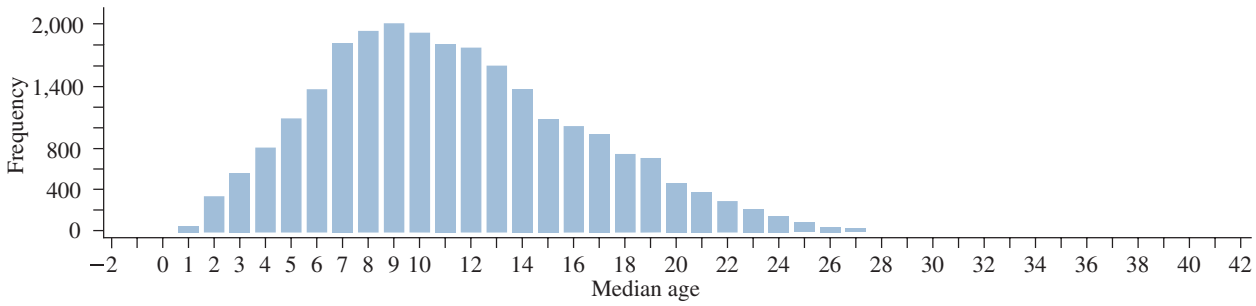
FIGURE 4.21 Sampling distribution of median for $n = 5, 10, 25$



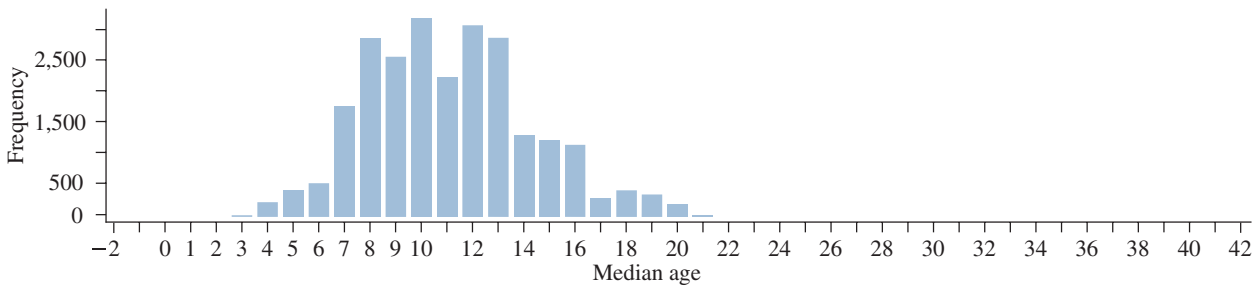
(a) Histogram of ages for 500 pennies



(b) Sampling distribution of median for $n = 5$

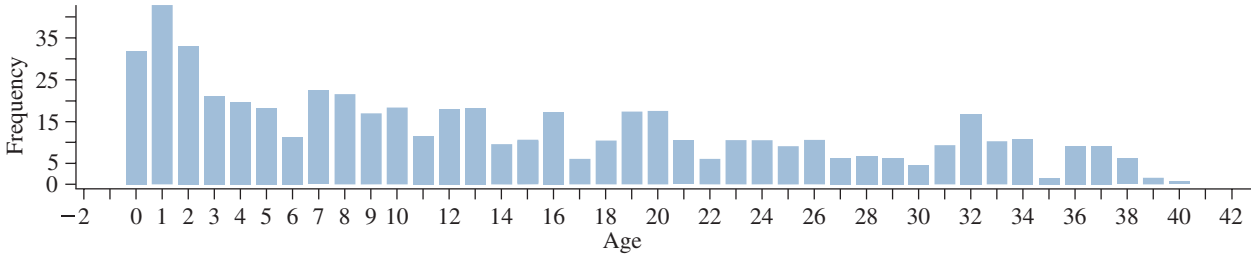


(c) Sampling distribution of median for $n = 10$

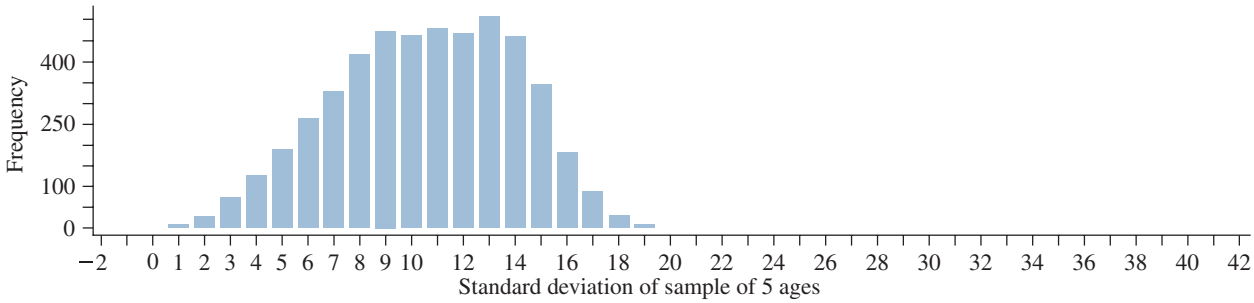


(d) Sampling distribution of median for $n = 25$

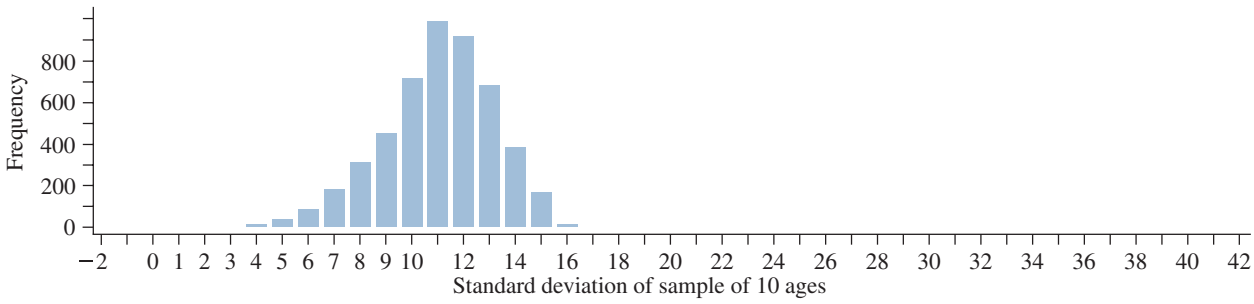
FIGURE 4.22 Sampling distribution of standard deviation for $n = 5, 10, 25$



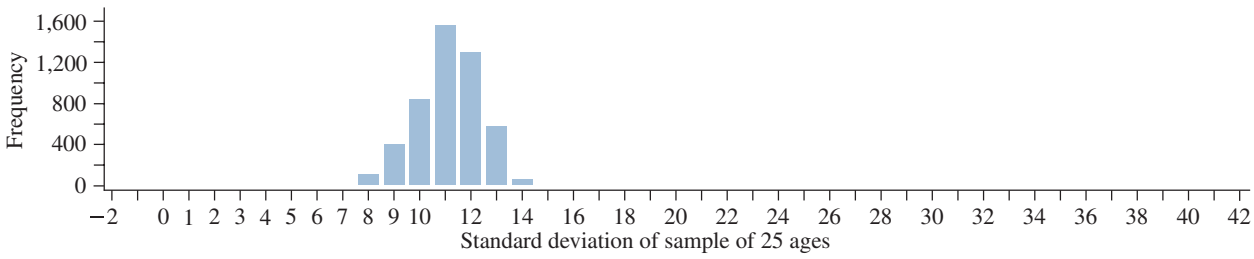
(a) Histogram of ages for 500 pennies



(b) Sampling distribution of standard deviation for $n = 5$



(c) Sampling distribution of standard deviation for $n = 10$



(d) Sampling distribution of standard deviation for $n = 25$

THEOREM 4.1

Central Limit Theorem for $\bar{y}$

Let $\bar{y}$ denote the sample mean computed from a random sample of n measurements from a population having a mean μ and finite standard deviation σ. Let $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$ denote the mean and standard deviation of the sampling distribution of $\bar{y}$, respectively. Based on repeated random samples of size n from the population, we can conclude the following:

1. $\mu_{\bar{y}} = \mu$
2. $\sigma_{\bar{y}} = \sigma/\sqrt{n}$
3. When n is large, the sampling distribution of $\bar{y}$ will be approximately normal (with the approximation becoming more precise as n increases).
4. When the population distribution is normal, the sampling distribution of $\bar{y}$ is exactly normal for any sample size n.

even for a very small sample size, $n = 10$, the shape of the sampling distribution of $\bar{y}$ is very similar to that of a normal distribution. This is not true in general. If the population distribution had many extreme values or several modes, the sampling distribution of $\bar{y}$ would require n to be considerably larger in order to achieve a symmetric bell shape.

We have seen that the sample size n has an effect on the shape of the sampling distribution of $\bar{y}$. The shape of the distribution of the population measurements also will affect the shape of the sampling distribution of $\bar{y}$. Figures 4.23 and 4.24 illustrate the effect of the population shape on the shape of the sampling distribution of $\bar{y}$. In Figure 4.23, the population measurements have a normal distribution. The sampling distribution of $\bar{y}$ is *exactly* a normal distribution for all values of n, as is illustrated for $n = 5$, 10, and 25 in Figure 4.23. When the population distribution is nonnormal, as depicted in Figure 4.24, the sampling distribution of $\bar{y}$ will not have a normal shape for small n (see Figure 4.24 with $n = 5$). However, for $n = 10$ and 25, the sampling distributions are nearly normal in shape, as can be seen in Figure 4.24.

It is very unlikely that the exact shape of the population distribution will be known. Thus, the exact shape of the sampling distribution of $\bar{y}$ will not be known either. The important point to remember is that the sampling distribution of $\bar{y}$ will

FIGURE 4.23
Sampling distribution of $\bar{y}$ for $n = 5, 10, 25$ when sampling from a normal distribution

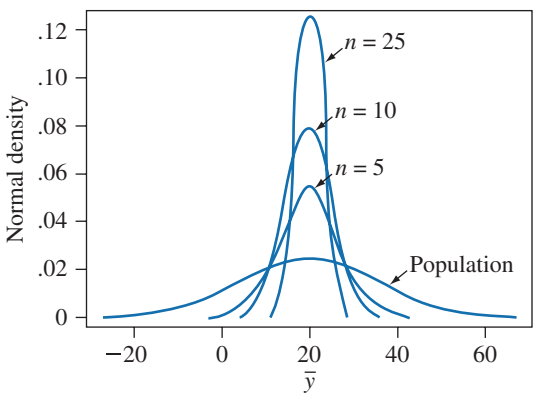
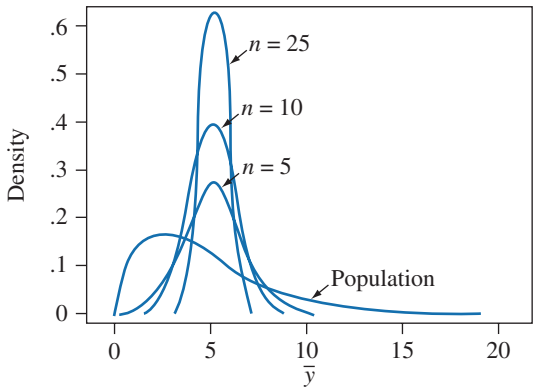


FIGURE 4.24
Sampling distribution of $\bar{y}$ for $n = 5, 10, 25$ when sampling from a skewed distribution



be approximately normally distributed with a mean $\mu_{\bar{y}} = \mu$, the population mean, and a standard deviation $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. The approximation will be more precise as n, the sample size for each sample, increases and as the shape of the population distribution becomes more like the shape of a normal distribution.

An obvious question is, How large should the sample size be for the Central Limit Theorem to hold? Numerous simulation studies have been conducted over the years, and the results of these studies suggest that, in general, the Central Limit Theorem holds for $n > 30$. However, one should not apply this rule blindly. If the population is heavily skewed, the sampling distribution for $\bar{y}$ will still be skewed even for $n > 30$. On the other hand, if the population is symmetric, the Central Limit Theorem holds for $n < 30$.

Therefore, take a look at the data. If the sample histogram is clearly skewed, then the population will also probably be skewed. Consequently, a value of n much higher than 30 may be required to have the sampling distribution of $\bar{y}$ be approximately normal. Any inference based on the normality of $\bar{y}$ for $n \leq 30$ under this condition should be examined carefully.

EXAMPLE 4.24

A person visits her doctor with concerns about her blood pressure. If the systolic blood pressure exceeds 150, the patient is considered to have high blood pressure and medication may be prescribed. A patient's blood pressure readings often have a considerable variation during a given day. Suppose a patient's systolic blood pressure readings during a given day have a normal distribution with a mean $\mu = 160$ mm mercury and a standard deviation $\sigma = 20$ mm.

a. What is the probability that a single blood pressure measurement will fail to detect that the patient has high blood pressure?
b. If five blood pressure measurements are taken at various times during the day, what is the probability that the average of the five measurements will be less than 150 and hence fail to indicate that the patient has high blood pressure?
c. How many measurements would be required in a given day so that there is at most a 1% probability of failing to detect that the patient has high blood pressure?

Solution Let y be the blood pressure measurement of the patient. y has a normal distribution with $\mu = 160$ and $\sigma = 20$.

a. $P(\text{measurement fails to detect high pressure}) = P(y \le 150) = P\left(z \le \frac{150 - 160}{20}\right) = P(z \le -0.5) = .3085$. Thus, there is over a 30% chance of failing to detect that the patient has high blood pressure if only a single measurement is taken.

b. Let $\bar{y}$ be the average blood pressure of the five measurements. Then $\bar{y}$ has a normal distribution with $\mu = 160$ and $\sigma = 20/\sqrt{5} = 8.944$.

$$P(\bar{y} \le 150) = P\left(z \le \frac{150 - 160}{8.944}\right) = P(z \le -1.12) = .1314$$

Therefore, by using the average of five measurements, the chance of failing to detect the patient has high blood pressure has been reduced from over 30% to about 13%.

c. We need to determine the sample size n such that $P(\bar{y} < 150) \le .01$. Now $P(\bar{y} < 150) = P\left(z \le \frac{150 - 160}{20/\sqrt{n}}\right)$. From the normal tables, we have $P(z \le -2.326) = .01$; therefore, $\frac{150 - 160}{20/\sqrt{n}} = -2.326$. Solving for n yields $\sqrt{n} = \frac{(-2.326)(20)}{-10}$, $n = 21.64$. It would require at least 22 measurements in order to achieve the goal of at most a 1% chance of failing to detect high blood pressure. ■

As demonstrated in Figures 4.21 and 4.22, the Central Limit Theorem can be extended to many different sample statistics. The form of the Central Limit Theorem for the sample median and sample standard deviation is considerably more complex than for the sample mean. Many of the statistics that we will encounter in later chapters will be either averages or sums of variables. The Central Limit Theorem for sums can be easily obtained from the Central Limit Theorem for the sample mean. Suppose we have a random sample of n measurements, $y_1, \ldots, y_n$, from a population and we let $\Sigma y = y_1 + \cdots + y_n$.

THEOREM 4.2

Central Limit Theorem for Σy

Let Σy denote the sum of a random sample of n measurements from a population having a mean μ and finite standard deviation σ. Let $\mu_{\Sigma y}$ and $\sigma_{\Sigma y}$ denote the mean and standard deviation of the sampling distribution of Σy, respectively. Based on repeated random samples of size n from the population, we can conclude the following:

1. $\mu_{\Sigma y} = n\mu$
2. $\sigma_{\Sigma y} = \sqrt{n}\sigma$
3. When n is large, the sampling distribution of Σy will be approximately normal (with the approximation becoming more precise as n increases).
4. When the population distribution is normal, the sampling distribution of Σy is exactly normal for any sample size n.

Usually, a sample statistic is used as an estimate of a population parameter. For example, a sample mean $\bar{y}$ can be used to estimate the population mean μ from which the sample was selected. Similarly, a sample median and sample standard deviation estimate the corresponding population median and standard deviation. The sampling distribution of a sample statistic is then used to determine how accurate

the estimate is likely to be. In Example 4.22, the population mean μ is known to be 6.5. Obviously, we do not know μ in any practical study or experiment. However, we can use the sampling distribution of $\bar{y}$ to determine the probability that the value of $\bar{y}$ for a random sample of $n = 2$ measurements from the population will be more than three units from μ. Using the data in Example 4.22, this probability is

$$P(2.5) + P(3) + P(10) + P(10.5) = \frac{4}{45}$$

In general, we would use the normal approximation from the Central Limit Theorem in making this calculation because the sampling distribution of a sample statistic is seldom known. This type of calculation will be developed in Chapter 5. Since a sample statistic is used to make inferences about a population parameter, the sampling distribution of the statistic is crucial in determining the accuracy of the inference.

interpretations of a sampling distribution

Sampling distributions can be **interpreted** in at least two ways. One way uses the long-run relative frequency approach. Imagine taking repeated samples of a fixed size from a given population and calculating the value of the sample statistic for each sample. In the long run, the relative frequencies for the possible values of the sample statistic will approach the corresponding sampling distribution probabilities. For example, if one took a large number of samples from the population distribution corresponding to the probabilities of Example 4.22 and, for each sample, computed the sample mean, approximately 9% would have $\bar{y} = 5.5$.

The other way to interpret a sampling distribution makes use of the classical interpretation of probability. Imagine listing all possible samples that could be drawn from a given population. The probability that a sample statistic will have a particular value (say, $\bar{y} = 5.5$) is then the proportion of all possible samples that yield that value. In Example 4.22, $P(\bar{y} = 5.5) = 4/45$ corresponds to the fact that 4 of the 45 samples have a sample mean equal to 5.5. Both the repeated-sampling and the classical method approaches to finding probabilities for a sample statistic are legitimate.

In practice, though, a sample is taken only once, and only one value of the sample statistic is calculated. A sampling distribution is not something you can see in practice; it is not an empirically observed distribution. Rather, it is a theoretical concept, a set of probabilities derived from assumptions about the population and about the sampling method.

There's an unfortunate similarity between the phrase "sampling distribution," meaning the theoretically derived probability distribution of a statistic, and the phrase "sample distribution," which refers to the histogram of individual values actually observed in a particular sample. The two phrases mean very different things. To avoid confusion, we will refer to the distribution of sample values as the **sample histogram** rather than as the sample distribution.

sample histogram

4.13 Normal Approximation to the Binomial

A binomial random variable y was defined earlier to be the number of successes observed in n independent trials of a random experiment in which each trial resulted in either a success (S) or a failure (F) and $P(S) = \pi$ for all n trials. We will now demonstrate how the Central Limit Theorem for sums enables us to calculate probabilities for a binomial random variable by using an appropriate normal curve as an approximation to the binomial distribution. We said in Section 4.8 that probabilities associated with values of y can be computed for a binomial experiment for any values of n or π, but the task becomes more difficult when n gets large. For example, suppose a sample of 1,000 voters is polled to determine sentiment toward

the consolidation of city and county government. What would be the probability of observing 460 or fewer favoring consolidation if we assume that 50% of the entire population favors the change? Here we have a binomial experiment with $n = 1{,}000$ and π, the probability of selecting a person favoring consolidation, equal to .5. To determine the probability of observing 460 or fewer favoring consolidation in the random sample of 1,000 voters, we could compute $P(y)$ using the binomial formula for $y = 460, 459, \ldots, 0$. The desired probability would then be

$$P(y = 460) + P(y = 459) + \cdots + P(y = 0)$$

There would be 461 probabilities to calculate, with each one being somewhat difficult because of the factorials. For example, the probability of observing 460 favoring consolidation is

$$P(y = 460) = \frac{1{,}000!}{460!540!}(.5)^{460}(.5)^{540}$$

A similar calculation would be needed for all other values of y.

To justify the use of the Central Limit Theorem, we need to define n random variables, $I_1, \ldots, I_n$, by

$$I_i = \begin{cases} 1 & \text{if the } i\text{th trial results in a success} \\ 0 & \text{if the } i\text{th trial results in a failure} \end{cases}$$

The binomial random variable y is the number of successes in the n trials. Now, consider the sum of the random variables $I_1, \ldots, I_n$: $\sum_{i=1}^{n} I_i$. A 1 is placed in the sum for each S that occurs and a 0 for each F that occurs. Thus, $\sum_{i=1}^{n} I_i$ is the number of Ss that occurred during the n trials. Hence, we conclude that $y = \sum_{i=1}^{n} I_i$. Because the binomial random variable y is the sum of independent random variables, each having the same distribution, we can apply the Central Limit Theorem for sums to y. Thus, the normal distribution can be used to approximate the binomial distribution when n is of an appropriate size. The normal distribution that will be used has a mean and standard deviation given by the following formulas:

$$\mu = n\pi, \quad \sigma = \sqrt{n\pi(1 - \pi)}$$

These are the mean and standard deviation of the binomial random variable y.

EXAMPLE 4.25

Use the normal approximation to the binomial to compute the probability of observing 460 or fewer favoring consolidation in a sample of 1,000 if we assume that 50% of the entire population favors the change.

Solution The normal distribution used to approximate the binomial distribution will have

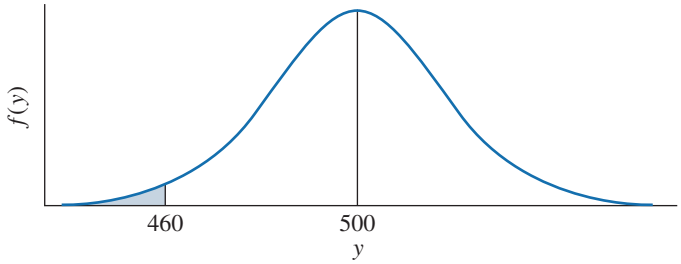
$$\mu = n\pi = 1{,}000(.5) = 500$$

$$\sigma = \sqrt{n\pi(1 - \pi)} = \sqrt{1{,}000(.5)(.5)} = 15.8$$

The desired probability is represented by the shaded area shown in Figure 4.25. We calculate the desired area by first computing

$$z = \frac{y - \mu}{\sigma} = \frac{460 - 500}{15.8} = -2.53$$

FIGURE 4.25
Approximating normal distribution for the binomial distribution, $\mu = 500$ and $\sigma = 15.8$



Referring to Table 1 in the Appendix, we find that the area under the normal curve to the left of 460 (for $z = -2.53$) is .0057. Thus, the probability of observing 460 or fewer favoring consolidation is approximately .0057. Using R, the exact value is **pbinom(460, 1000, .5)** = .0062. ■

The normal approximation to the binomial distribution can be unsatisfactory if $n\pi < 5$ or $n(1 - \pi) < 5$. If π, the probability of success, is small and n, the sample size, is modest, the actual binomial distribution is seriously skewed to the right. In such a case, the symmetric normal curve will give an unsatisfactory approximation. If π is near 1, so $n(1 - \pi) < 5$, the actual binomial will be skewed to the left, and, again, the normal approximation will not be very accurate. The normal approximation, as described, is quite good when $n\pi$ and $n(1 - \pi)$ exceed about 20. In the middle zone, $n\pi$ or $n(1 - \pi)$ between 5 and 20, a modification called a **continuity correction** makes a substantial contribution to the quality of the approximation.

continuity correction

The point of the continuity correction is that we are using the continuous normal curve to approximate a discrete binomial distribution. A picture of the situation is shown in Figure 4.26.

The binomial probability that $y \leq 5$ is the sum of the areas of the rectangles above 5, 4, 3, 2, 1, and 0. This probability (area) is approximated by the area under the superimposed normal curve to the left of 5. Thus, the normal approximation ignores half of the rectangle above 5. The continuity correction simply includes the area between $y = 5$ and $y = 5.5$. For the binomial distribution with $n = 20$ and $\pi = .30$ (pictured in Figure 4.26), the correction is to take $P(y \leq 5)$ as $P(y \leq 5.5)$. Instead of

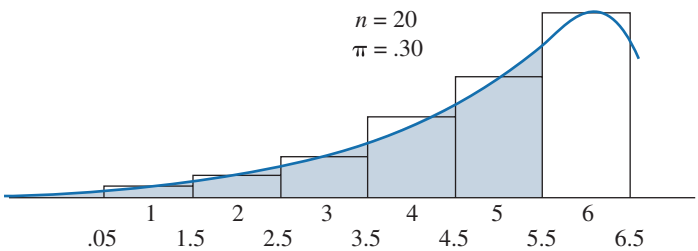
$$P(y \leq 5) = P[z \leq (5 - 20(.3))/\sqrt{20(.3)(.7)}] = P(z \leq -.49) = .3121$$

use

$$P(y \leq 5.5) = P[z \leq (5.5 - 20(.3))/\sqrt{20(.3)(.7)}] = P(z \leq -.24) = .4052$$

The actual binomial probability is **pbinom(5, 20, .3)** = .4164. The general idea of the continuity correction is to add or subtract .5 from a binomial value before using normal probabilities. The best way to determine whether to add or subtract is to draw a picture like Figure 4.26.

FIGURE 4.26
Normal approximation to the binomial



Normal Approximation to the Binomial Probability Distribution

For large n and π not too near 0 or 1, the distribution of a binomial random variable y may be approximated by a normal distribution with $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1-\pi)}$. This approximation should be used only if $n\pi \geq 5$ and $n(1-\pi) \geq 5$. A continuity correction will improve the quality of the approximation in cases in which n is not overwhelmingly large.

EXAMPLE 4.26

A large drug company has 100 potential new prescription drugs under clinical test. About 20% of all drugs that reach this stage are eventually licensed for sale. What is the probability that at least 15 of the 100 drugs are eventually licensed? Assume that the binomial assumptions are satisfied, and use a normal approximation with continuity correction.

Solution Let y be the number of approved drugs. We are assuming y has a binomial distribution with $n = 100$ and $\pi = .2$. The mean of y is $\mu = 100(.2) = 20$, and the standard deviation is $\sqrt{100(.2)(.8)} = 4$. Because $n\pi = 100(.2) = 20 > 5$ and $n(1-\pi) = 100(.8) = 80 > 5$, the normal approximation can safely be used to approximate the probability that 15 or more drugs are approved; that is, $P(y \geq 15)$. Because $y = 15$ is included, the continuity correction is to take the event as y greater than or equal to 14.5.

$$P(y \geq 15) \approx P\left(z \geq \frac{14.5 - 20}{4}\right) = P(z \geq -1.375) = 1 - P(z < -1.375)$$
$$= 1 - .0846 = .9154$$

Using the R command for computing binomial probabilities, the exact probability is $P(y \geq 15) = 1 - P(y \leq 14) = 1 -$ **pbinom (14, 100, .2)** $= .9196$. Comparing the approximate probability, .9154, to the exact probability, .9196, we can conclude that the approximation was accurate to two decimal places.

If the continuity correction was not used, the probability would be approximated to be

$$P(y \geq 15) \approx P\left(z \geq \frac{15 - 20}{4}\right) = P(z \geq -1.25) = 1 - P(z < -1.25)$$
$$= 1 - .1056 = .8944$$

Thus, the continuity correction is crucial in obtaining an accurate approximation. ■

4.14 Evaluating Whether or Not a Population Distribution Is Normal

In many scientific experiments or business studies, the researcher wishes to determine if a normal distribution would provide an adequate fit to the population distribution. This would allow the researcher to make probability calculations and draw inferences about the population based on a random sample of observations from that population. Knowledge that the population distribution is not normal also may provide the researcher insight concerning the population under study. This may indicate that the physical mechanism generating the data has been altered or is of a form different from previous specifications. Many of the statistical

procedures that will be discussed in subsequent chapters of this book require that the population distribution have a normal distribution or at least be adequately approximated by a normal distribution. In this section, we will provide a graphical procedure and a quantitative assessment of how well a normal distribution models the population distribution.

normal probability plot

The graphical procedure that will be constructed to assess whether a random sample $y_1, y_2, \ldots, y_n$ was selected from a normal distribution is referred to as a **normal probability plot** of the data values. This plot is a variation on the quantile plot that was introduced in Chapter 3. In the normal probability plot, we compare the quantiles from the data observed from the population to the corresponding quantiles from the standard normal distribution. Recall that the quantiles from the data are just the data ordered from smallest to largest: $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$, where $y_{(1)}$ is the smallest value in the data $y_1, y_2, \ldots, y_n$; $y_{(2)}$ is the second smallest value; and so on until reaching $y_{(n)}$, which is the largest value in the data. Sample quantiles separate the sample in the same fashion as the population percentiles, which were defined in Section 4.10. Thus, the sample quantile $Q(u)$ has at least $100u\%$ of the data values less than $Q(u)$ and has at least $100(1-u)\%$ of the data values greater than $Q(u)$. For example, $Q(.1)$ has at least 10% of the data values less than $Q(.1)$ and has at least 90% of the data values greater than $Q(.1)$. $Q(.5)$ has at least 50% of the data values less than $Q(.5)$ and has at least 50% of the data values greater than $Q(.5)$. Finally, $Q(.75)$ has at least 75% of the data values less than $Q(.75)$ and has at least 25% of the data values greater than $Q(.75)$. This motivates the following definition for the sample quantiles.

DEFINITION 4.14

Let $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ be the ordered values from a data set. The $[(i-.5)/n]th$ sample quantile, $Q((i-.5)/n)$, is $y_{(i)}$. That is, $y_{(1)} = Q((.5)/n)$ is the $[(.5)/n]th$ sample quantile, $y_{(2)} = Q((1.5)/n)$ is the $[(1.5)/n]th$ sample quantile,…,and, lastly, $y_{(n)} = Q((n-.5)/n)$ is the $[(n-.5)/n]th$ sample quantile.

Suppose we had a sample of $n = 20$ observations: $y_1, y_2, \ldots, y_{20}$. Then

$y_{(1)} = Q((.5)/20) = Q(.025)$ is the $.025th$ sample quantile,
$y_{(2)} = Q((1.5)/20) = Q(.075)$ is the $.075th$ sample quantile,
$y_{(3)} = Q((2.5)/20) = Q(.125)$ is the $.125th$ sample quantile,…, and
$y_{(20)} = Q((19.5)/20) = Q(.975)$ is the $.975th$ sample quantile.

In order to evaluate whether a population distribution is normal, a random sample of n observations is obtained, the sample quantiles are computed, and these n quantiles are compared to the corresponding quantiles computed using the conjectured population distribution. If the conjectured distribution is the normal distribution, then we would use the normal tables to obtain the quantiles $z_{(i-.5)/n}$ for $i = 1, 2, \ldots, n$. The normal quantiles are obtained from the standard normal tables, Table 1 in the Appendix, for the n values $.5/n, 1.5/n, \ldots, (n-.5)/n$. For example, if we had $n = 20$ data values, then we would obtain the normal quantiles for $.5/20 = .025$, $1.5/20 = .075$, $2.5/20 = .125, \ldots, (20-.5)/20 = .975$. From Table 1, we find that these quantiles are given by $z_{.025} = -1.960$, $z_{.075} = -1.440$, $z_{.125} = -1.150, \ldots$, $z_{.975} = 1.960$. The normal quantile plot is obtained by plotting the n pairs of points:

$$(z_{.5/n}, y_{(1)});\ (z_{1.5/n}, y_{(2)});\ (z_{2.5/n}, y_{(3)}); \ldots ; (z_{(n-.5)/n}, y_{(n)})$$

If the population from which the sample of n values was randomly selected has a normal distribution, then the plotted points should fall close to a straight line. The following example will illustrate these ideas.

EXAMPLE 4.27

It is generally assumed that cholesterol readings in large populations have a normal distribution. In order to evaluate this conjecture, the cholesterol readings of $n = 20$ patients were obtained. These are given in Table 4.12, along with the corresponding normal quantile values. It is important to note that the cholesterol readings are given in an ordered fashion from smallest to largest. The smallest cholesterol reading is matched with the smallest normal quantile, the second-smallest cholesterol reading with the second-smallest quantile, and so on. Obtain the normal quantile plot for the cholesterol data, and assess whether the data were selected from a population having a normal distribution.

Solution

TABLE 4.12
Sample and normal quantiles for cholesterol readings

Patient	Cholesterol Reading	$(i - .5)/20$	Normal Quantile
1	133	.025	−1.960
2	137	.075	−1.440
3	148	.125	−1.150
4	149	.175	−.935
5	152	.225	−.755
6	167	.275	−.598
7	174	.325	−.454
8	179	.375	−.319
9	189	.425	−.189
10	192	.475	−.063
11	201	.525	.063
12	209	.575	.189
13	210	.625	.319
14	211	.675	.454
15	218	.725	.598
16	238	.775	.755
17	245	.825	.935
18	248	.875	1.150
19	253	.925	1.440
20	257	.975	1.960

A plot of the sample quantiles versus the corresponding normal quantiles is displayed in Figure 4.27. The plotted points generally follow a straight-line pattern.

FIGURE 4.27
Normal quantile plot for cholesterol reading

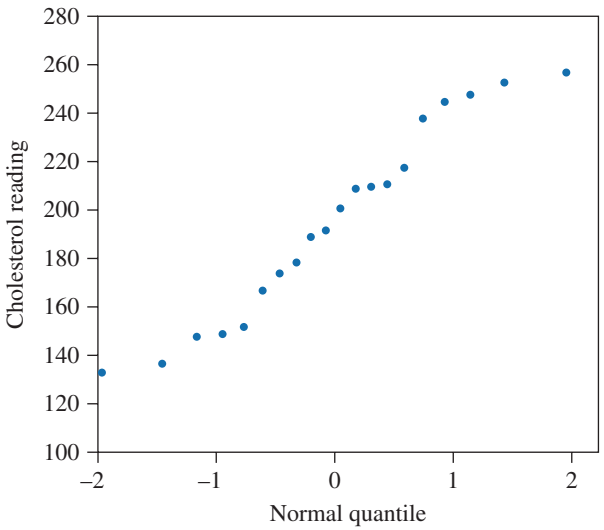
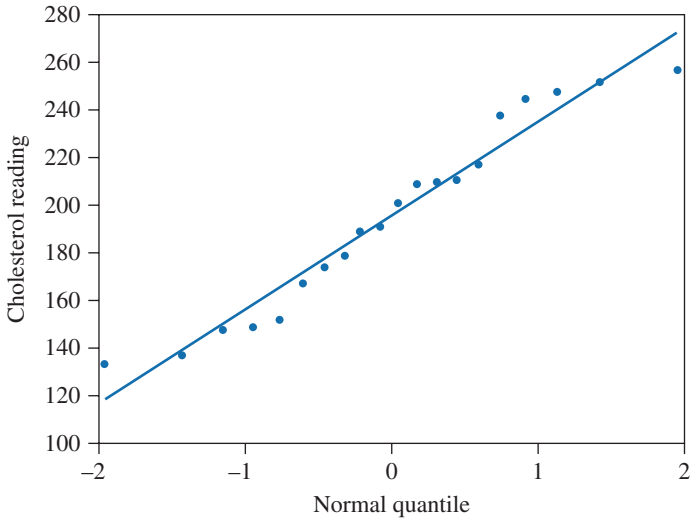


FIGURE 4.28
Normal quantile plot for cholesterol reading



Using the R code in Section 4.16, we can obtain a plot with a fitted line that assists us in assessing how close the plotted points fall relative to a straight line. This plot is displayed in Figure 4.28. The 20 points appear to be relatively close to the fitted line, and, thus, the normal quantile plot would appear to suggest that the normality of the population distribution is plausible.

Using a graphical procedure, there is a high degree of subjectivity in making an assessment of how well the plotted points fit a straight line. The scales of the axes on the plot can be increased or decreased, resulting in a change in our assessment of fit. Therefore, a quantitative assessment of the degree to which the plotted points fall near a straight line will be introduced.

In Chapter 3, we introduced the sample correlation coefficient r to measure the degree to which two variables satisfied a linear relationship. We will now discuss how this coefficient can be used to assess our certainty that the sample data were selected from a population having a normal distribution. First, we must alter which normal quantiles are associated with the ordered data values. In the above discussion, we used the normal quantiles corresponding to $(i - .5)/n$. In calculating the correlation between the ordered data values and the normal quantiles, a more precise measure is obtained if we associate the $(i - .375)/(n + .25)$ normal quantiles for $i = 1, \ldots, n$ with the n data values $y_{(1)}, \ldots, y_{(n)}$. We then calculate the value of the correlation coefficient, r, from the n pairs of values. To provide a more definitive assessment of our level of certainty that the data were sampled from a normal distribution, we then obtain a value from Table 15 in the Appendix. This value, called a p-value, can then be used along with the following criterion (Table 4.13) to rate the degree of fit of the data to a normal distribution.

TABLE 4.13
Criteria for assessing fit of normal distribution

p-value	Assessment of Normality
$p < .01$	Very poor fit
$.01 \le p < .05$	Poor fit
$.05 \le p < .10$	Acceptable fit
$.10 \le p < .50$	Good fit
$p \ge .50$	Excellent fit

It is very important that the normal quantile plot accompany the calculation of the correlation because large sample sizes may result in an assessment of a poor fit when the graph would indicate otherwise. The following example will illustrate the calculations involved in obtaining the correlation.

EXAMPLE 4.28

Consider the cholesterol data in Example 4.27. Calculate the correlation coefficient, and make a determination of the degree of fit of the data to a normal distribution.

Solution The data are summarized in Table 4.14 along with their corresponding normal quantiles.

TABLE 4.14
Normal quantiles data

Patient	Cholesterol Reading	$(i - .375)/(20 + .25)$	Normal Quantile
i	y_i		x_i
1	133	.031	−1.868
2	137	.080	−1.403
3	148	.130	−1.128
4	149	.179	−.919
5	152	.228	−.744
6	167	.278	−.589
7	174	.327	−.448
8	179	.377	−.315
9	189	.426	−.187
10	192	.475	−.062
11	201	.525	.062
12	209	.574	.187
13	210	.623	.315
14	211	.673	.448
15	218	.722	.589
16	238	.772	.744
17	245	.821	.919
18	248	.870	1.128
19	253	.920	1.403
20	257	.969	1.868

The calculation of the correlation between cholesterol reading (y) and normal quantile (x) will be done in Table 4.15. First, we compute $\bar{y} = 195.5$ and $\bar{x} = 0$. Then the calculation of the correlation will proceed as in our calculations from Chapter 3.

The correlation is then computed as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}} = \frac{720.18}{\sqrt{(17.634)(30511)}} = .982$$

From Table 15 in the Appendix with $n = 20$ and $r = .982$, we obtain p-value $\approx .50$. This value is obtained by locating the number in the row for $n = 20$ that is closest to $r = .982$. The α-value heading this column is the p-value. Thus, we would appear to have an excellent fit between the sample data and the normal distribution. This is consistent with the fit that is displayed in Figure 4.28, where the 20 plotted points are very near to the straight line. The R command **cor(*y*, *x*)** yields the value .9818, where y and x are the values in Table 4.14.

TABLE 4.15
Calculation of correlation coefficient

$(x_i - \bar{x})$ $(x_i - 0)$	$(y_i - \bar{y})$ $(y_i - 195.5)$	$(x_i - \bar{x})(y_i - \bar{y})$ $(x_i - 0)(y_i - 195.5)$	$(y_i - \bar{y})^2$ $(y_i - 195.5)^2$	$(x_i - \bar{x})^2$ $(x_i - 0)^2$
−1.868	−62.5	116.765	3,906.25	3.49033
−1.403	−58.5	82.100	3,422.25	1.96957
−1.128	−47.5	53.587	2,256.25	1.27271
−.919	−46.5	42.740	2,162.25	.84481
−.744	−43.5	32.370	1,892.25	.55375
−.589	−28.5	16.799	812.25	.34746
−.448	−21.5	9.627	462.25	.20050
−.315	−16.5	5.190	272.25	.09896
−.187	−6.5	1.214	42.25	.03488
−.062	−3.5	.217	12.25	.00384
.062	5.5	.341	30.25	.00384
.187	13.5	2.521	182.25	.03488
.315	14.5	4.561	210.25	.09896
.448	15.5	6.940	240.25	.20050
.589	22.5	13.263	506.25	.34746
.744	42.5	31.626	1,806.25	.55375
.919	49.5	45.497	2,450.25	.84481
1.128	52.5	59.228	2,756.25	1.27271
1.403	57.5	80.696	3,306.25	1.96957
1.868	61.5	114.897	3,782.25	3.49033
0	0	720.18	30,511	17.634

4.15 RESEARCH STUDY: Inferences About Performance-Enhancing Drugs Among Athletes

As was discussed in the abstract to the research study given at the beginning of this chapter, the use of performance-enhancing substances has two major consequences: the artificial enhancement of performance (known as doping) and the use of potentially harmful substances that may have significant health effects for the athlete. However, failing a drug test can devastate an athlete's career. The controversy over performance-enhancing drugs has seriously brought into question the reliability of the tests for these drugs. The article in ***Chance*** discussed at the beginning of this chapter examines the case of Olympic runner Mary Decker Slaney. Ms. Slaney was a world-class distance runner during the 1970s and 1980s. After a series of illnesses and injuries, she was forced to stop competitive running. However, at the age of 37, Slaney made a comeback in long-distance running. Slaney submitted to a mandatory test of her urine at the 1996 U.S. Olympic Trials. The results indicated that she had elevated levels of testosterone and hence may have used a banned performance-enhancing drug. Her attempt at a comeback was halted by her subsequent suspension by USA Track and Field (USATF). Slaney maintained her innocence throughout a series of hearings before USATF and was exonerated in September 1997 by a Doping Hearing Board of the USATF. However, the U.S. Olympic Committee (USOC) overruled the USATF decision and stated that Slaney was guilty of a doping offense. Although Slaney continued to maintain that she had never used the drug, her career as a competitive runner was terminated. Anti-doping officials regard a positive test result as irrefutable evidence that an illegal drug was used, to the exclusion of any other explanation. We will now address how the use of Bayes' Formula, the sensitivity and specificity

of a test, and the prior probability of drug use can be used to explain to anti-doping officials that drug tests can be wrong.

We will use tests for detecting artificial increases in testosterone concentrations to illustrate the various concepts involved in determining the reliability of a testing procedure. The article states, "Scientists have attempted to detect artificial increases in testosterone concentrations through the establishment of a 'normal urinary range' for the T/E ratio." Despite the many limitations in setting this limit, scientists set the threshold for positive testosterone doping at a T/E ratio greater than 6:1. The problem is to determine the probabilities associated with various tests for the T/E ratio. In particular, what is the probability that an athlete is a banned-drug user given she tests positive for the drug (positive predictive value, or PPV)?

We will use the example given in the article. Suppose in a population of 1,000 athletes there are 20 users. That is, prior to testing a randomly selected athlete for the drug, there is a 20/1,000 = 2% chance that the athlete is a user (the prior probability of randomly selecting a user is .02 = 2%). Suppose the testing procedure has a sensitivity of 80% and a specificity of 99%. Thus, 16 of the 20 users would test positive, 20(.8) = 16, and about 10 of the nonusers would test positive, 980(1 − .99) = 9.8. If an athlete tests positive, what is the probability she is a user? We now have to make use of Bayes' Formula to compute PPV.

$$\text{PPV} = \frac{\text{sens} * \text{prior}}{\text{sens} * \text{prior} + (1 - \text{spec}) * (1 - \text{prior})}$$

where "sens" is the sensitivity of the test, "spec" is the specificity of the test, and "prior" is the prior probability that an athlete is a banned-drug user. For our example with a population of 1,000 athletes,

$$\text{PPV} = \frac{(.8) * (20/1{,}000)}{(.8) * (20/1{,}000) + (1 - .99) * (1 - 20/1{,}000)} = .62$$

Therefore, if an athlete tests positive, there is only a 62% chance that she has used the drug. Even if the sensitivity of the test is increased to 100%, the PPV is still relatively small:

$$\text{PPV} = \frac{(1) * (20/1{,}000)}{(1) * (20/1{,}000) + (1 - .99) * (1 - 20/1{,}000)} = .67$$

There is a 33% chance that the athlete is a nonuser even though the test result was positive. Thus, if the prior probability is small, there will always be a high degree of uncertainty with the test result even when the test has values of sensitivity and specificity near 1.

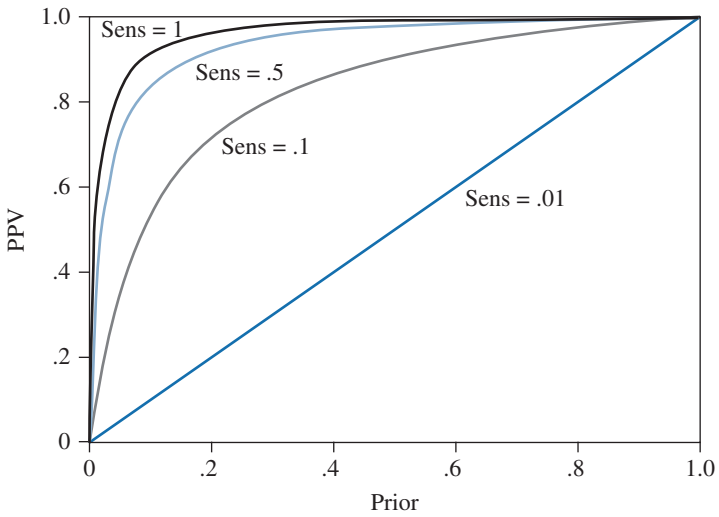
However, if the prior probability is fairly large, then the PPV will be much closer to 1. For example, if the population consists of 900 users and only 100 nonusers and if the testing procedure has sensitivity = .9 and specificity = .99, then the PPV would be .9988:

$$\text{PPV} = \frac{(.9) * (900/1{,}000)}{(.9) * (900/1{,}000) + (1 - .99) * (1 - 900/1{,}000)} = .9988$$

That is, the chance that the tested athlete is a user given she produced a positive test would be 99.88%, a very small chance of a false positive.

From this, we conclude that an essential factor in Bayes' Formula is the prior probability of an athlete being a banned-drug user. Making matters even worse in this situation is the fact that the prevalence (prior probability) of substance abuse

FIGURE 4.29
Relationship between PPV and prior probability for four different values of sensitivity; all curves assume specificity is 99%



is very difficult to determine. Hence, there will inevitably be a subjective aspect to assigning a prior probability. The authors of the article comment on the selection of the prior probability, suggesting that in their particular sport, a hearing board consisting of athletes participating in the same sport as the athlete being tested would be especially appropriate for making decisions about prior probabilities. For example, assuming the board knows nothing about the athlete beyond what is presented at the hearing, it might regard drug abuse as rare, and, hence, the PPV would be at most moderately large. On the other hand, if the board knew that drug abuse is widespread, then the probability of abuse would be larger, based on a positive test result.

To investigate further the relationship among PPV, prior probability, and sensitivity for a fixed specificity of 99%, consider Figure 4.29. The calculations of PPV are obtained by using Bayes' Formula for a selection of prior and sensitivity, and with specificity = .99.

We can thus observe that if the sensitivity of the test is relatively low—say, less than 50%—then unless the prior is above 20%, we will not be able to achieve a PPV greater than 90%. The article describes how the above figure allows for using Bayes' Formula in reverse. For example, a hearing board may make the decision that it would not rule against an athlete unless his or her probability of being a user was at least 95%. Suppose we have a test having both sensitivity and specificity of 99%. Then the prior probability must be at least 50% in order to achieve a PPV of 95%. This would allow the board to use its knowledge about the prevalence of drug abuse in the population of athletes to determine if a prevalence of 50% or larger is realistic.

The authors conclude with the following comments:

> Conclusions about the likelihood of testosterone doping require consideration of three components: specificity and sensitivity of the testing procedure, and the prior probability of use. As regards the T/E ratio, anti-doping officials consider only specificity. The result is a flawed process of inference. Bayes' rule shows that it is impossible to draw conclusions about guilt on the basis of specificity alone. Policy-makers in the athletic federations should follow the lead of medical scientists who use sensitivity, specificity, and Bayes' rule in interpreting diagnostic evidence.

4.16 R Instructions

Generating Random Numbers

To generate 1,000 random numbers from the integers [0, 1,…, 9]:

1. $y = c(0{:}9)$
2. $x = \text{sample}(y, 1000, \text{replace}=\text{T})$
3. x

Calculating Binomial Probabilities

To calculate binomial probabilities when X has a binomial distribution with $n = 10$ and $\pi = 0.6$:

1. To calculate $P(X = 3)$, use the command **dbinom(3, 10, .6)**
2. To calculate $P(X \leq 3)$, use the command **pbinom(3, 10, .6)**
3. To calculate $P(X = k)$ for $k = 0, 1, \ldots, 10$, use the commands $k = \text{c}(0 : 10)$ and **dbinom(*k*, 10, .6)**

Calculating Poisson Probabilities

To calculate Poisson probabilities when Y has a binomial distribution with $\lambda = 10$ and $\pi = 0.6$:

1. To calculate $P(X = 3)$, use the command **dbinom(3, 10, .6)**
2. To calculate $P(X \leq 3)$, use the command **pbinom(3, 10, .6)**
3. To calculate $P(X = k)$ for $k = 0, 1, \ldots, 10$, use the commands $k = \text{c}(0 : 10)$ and **dbinom(*k*, 10, .6)**

Calculating Normal Probabilities

To calculate probabilities when X has a normal distribution with $\mu = 23$ and $\sigma = 5$:

1. To calculate $P(X \leq 18)$, use the command **pnorm(18, 23, 5)**
2. To calculate $P(X > 18)$, use the command **1 − pnorm(18, 23, 5)**
3. To find 85th percentile, use **q(.85, 23, 5)**

Generating Sampling Distribution of $\bar{y}$

The following R commands will simulate the sampling distribution of $\bar{y}$. We will generate 10,000 values of $\bar{y}$, with each of the 10,000 values of $\bar{y}$ computed from a unique random sample of 16 observations, from a population having a normal distribution with $\mu = 43$ and $\sigma = 7$.

1. $r = 10{,}000$
2. $y = \text{rep}(0, 16)$
3. $y\text{bar16} = \text{rep}(0, r)$
4. for (i in 1:r){
5. $y = \textbf{rnorm}(16, 43, 7)$
6. $y\text{bar16}[\text{i}] = \textbf{mean}(y)$ }

The above commands will produce 10,000 values for $\bar{y}$, where $\bar{y}$ is the average of 16 data values from a population having a normal distribution with $\mu = 43$ and $\sigma = 7$. To display the 10,000 values, type "ybar16".

The following three commands will generate a histogram, mean, and standard deviation for the 10,000 values:

1. **hist**(ybar16)
2. **mean**(ybar16)
3. **sd**(ybar16)

The histogram should be bell-shaped with its center near 43. The mean of the 10,000 values should be close to 43, and the standard deviation should be close to $7/\sqrt{16} = 1.75$.

Commands to Generate the Plot in Figure 4.28

The following R commands will generate the normal reference plot in Figure 4.28 and the correlation coefficient.

1. $y = c(133, 137, 148, 149, 152, 167, 174, 179, 189, 192, 201, 209, 210, 211, 218, 238, 245, 248, 253, 257)$
2. $y = \text{sort}(y)$
3. $n = \text{length}(y)$
4. $i = 1 : \text{n}$
5. $u = (i - .375)/(\text{n} + .25)$
6. $x =$ **qnorm**(u)
7. plot(x, y, xlab = "Normal quantiles", ylab = "Cholesterol readings", lab = $c(7, 8, 7)$, ylim = $c(100, 280)$, main = "Normal Reference Distribution Plot\n Cholesterol readings", cex = .95)
8. **abline**(lm(y ~ x))
9. **cor**(x, y)

4.17 Summary and Key Formulas

In this chapter, we presented an introduction to probability, probability distributions, and sampling distributions. Knowledge of the probabilities of sample outcomes is vital to a statistical inference. Three different interpretations of the probability of an outcome were given: the classical, relative frequency, and subjective interpretations. Although each has a place in statistics, the relative frequency approach has the most intuitive appeal because it can be checked.

Quantitative random variables are classified as either discrete or continuous random variables. The probability distribution for a discrete random variable y is a display of the probability $P(y)$ associated with each value of y. This display may be presented in the form of a histogram, table, or formula.

The binomial is a very important and useful discrete random variable. Many experiments that scientists conduct are similar to a coin-tossing experiment where dichotomous (yes–no) types of data are accumulated. The binomial experiment frequently provides an excellent model for computing probabilities of various sample outcomes.

Probabilities associated with a continuous random variable correspond to areas under the probability distribution. Computations of such probabilities were illustrated for areas under the normal curve. The importance of this exercise is borne out by the Central Limit Theorem: Any random variable that is expressed as a sum or average of a random sample from a population having a finite standard deviation will have a normal distribution for a sufficiently large sample size.

Direct application of the Central Limit Theorem gives the sampling distribution for the sample mean. Because many sample statistics are either sums or averages of random variables, application of the Central Limit Theorem provides us with information about probabilities of sample outcomes. These probabilities are vital for the statistical inferences we wish to make.

Key Formulas

1. Bayes' Formula

   If $A_1, A_2, \ldots, A_k$ are mutually exclusive events and B is any event, then

   $$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}$$

2. Binomial probability

   $$P(y = k) = \frac{n!}{k!(n-k)}\pi^k(1-\pi)^{n-k} = \boldsymbol{dbinom(k, n, \pi)} \text{ using R function}$$

   $$P(y \le k) = \sum_{i=0}^{k} P(y = i) = \boldsymbol{pbinom(k, n, \pi)} \text{ using R function}$$

3. Poisson probability

   $$P(y = k) = \frac{e^{-\mu}\mu^k}{k!} = \boldsymbol{dpois(k, \mu)} \text{ using R function}$$

   $$P(y \le k) = \sum_{i=0}^{k} P(y = i) = \boldsymbol{ppois(k, \mu)} \text{ using R function}$$

4. Normal probability

   Let y have a normal distribution with mean μ and standard deviation σ, and let z have a standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

   $$P(y \le w) = P\left(z \le \frac{w-\mu}{\sigma}\right) = \boldsymbol{pnorm}\left(\frac{\boldsymbol{w-\mu}}{\boldsymbol{\sigma}}\right) \text{ using R code}$$

5. Sampling distribution for sample mean $\bar{y}$ when random sample is from population having mean μ and standard deviation σ

   Mean: μ
   Standard deviation: $\sigma/\sqrt{n}$
   For a large sample size n, the distribution of $\bar{y}$ will be approximately a normal distribution.

6. Normal approximation to binomial distribution

   $$\mu = n\pi, \quad \sigma = \sqrt{n\pi(1-\pi)}$$

   Provided both $n\pi \ge 5$ and $n(1-\pi) \ge 5$,

   $$P(y \le k) \approx P\left(z \le \frac{k + .5 - \mu}{\sigma}\right) = \boldsymbol{pnorm}\left(\frac{\boldsymbol{k + .5 - \mu}}{\boldsymbol{\sigma}}\right)$$

   $$P(y \ge k) \approx P\left(z \ge \frac{k - .5 - \mu}{\sigma}\right) = 1 - \boldsymbol{pnorm}\left(\frac{\boldsymbol{k - .5 - \mu}}{\boldsymbol{\sigma}}\right)$$

   Compare the above to the exact values:

   $$P(y \le k) = \boldsymbol{pbinorm(k, n, \pi)}$$

   $$P(y \ge k) = 1 - P(y \le k-1) = 1 - \boldsymbol{pbinorm(k-1, n, \pi)}$$

4.18 Exercises

4.1 Introduction and Abstract of Research Study

Basic **4.1** Indicate which interpretation of the probability statement seems most appropriate.

- **a.** A casino in New Jersey posts a probability of .02 that the Dallas Cowboys will win Super Bowl L.
- **b.** A purchaser of a single ticket in the Texas Powerball has a probability of 1/175,223,510 of winning the big payout.
- **c.** The quality control engineer of a large pharmaceutical firm conducts an intensive process reliability study. Based on the findings of the study, the engineer claims that the probability that a bottle of a newly produced drug will have a shelf life greater than 2 years is .952.
- **d.** The probability that the control computer on a nuclear power plant and its backup will both fail is .00001.
- **e.** The state meteorologist of Michigan reports that there is a 70/30 chance that the rainfall during the months of June through August in 2014 will be below normal; that is, there is a .70 probability of the rainfall being below normal and a .30 probability of the rainfall being above normal.
- **f.** A miniature tablet that is small enough to be worn as a watch is in beta testing. In a preliminary report, the company states that more than 55% the 500 testers found the device to be easier to use than a full-sized tablet. The probability of this happening is .011 provided there is no difference in ease of use of the two devices.

Med. **4.2** If you are having a stroke, it is critical that you get medical attention right away. Immediate treatment may minimize the long-term effects of a stroke and even prevent death. A major U.S. city reported that there was a 1 in 250 chance of the patient not having long-term memory problems after suffering a stroke. That is, for a person suffering a stroke in the city, P(no memory problems) = 1/250 = .004. This very high chance of memory problems was attributed to many factors associated with large cities that affected response times, such as heavy traffic, the misidentification of addresses, and the use of cell phones, which results in emergency personnel not being able to obtain an address. The study documented the 1/250 probability based on a study of 15,000 requests for assistance by stroke victims.

- **a.** Provide a relative frequency interpretation of the .004 probability.
- **b.** The value .004 was based on the records of 15,000 requests for assistance from stroke victims. How many of the 15,000 victims in the study had long-term memory problems? Explain your answer.

Gov. **4.3** In reporting highway safety, the ***National Highway Traffic Safety Administration (NHTSA)*** reports the number of deaths in automobile accidents each year. If there is a decrease in the number of traffic deaths from the previous year, NHTSA claims that the chance of a death on the highways has decreased. Explain the flaw in NHTSA's claim.

Bus. **4.4** In a cable TV program concerning the risk of travel accidents, it was stated that the chance of a fatal airplane crash was 1 in 11 million. An explanation of this risk was that you could fly daily for the next 11 million days (30,137 years) before you would experience a fatal crash. Provide an explanation why this statement is misleading.

Game **4.5** The gaming commission in its annual examination of the casinos in the state reported that all roulette wheels were fair. Explain the meaning of the term *fair* with respect to the roulette wheel?

4.2 Finding the Probability of an Event

Edu. **4.6** Suppose an economics examination has 25 true-or-false questions and a passing grade is obtained with 17 or more correct answers. A student answers the 25 questions by flipping a fair coin and answering true if the coin shows a head and false if it shows a tail.

a. Using the classical interpretation of probability, what is the chance the student will pass the exam?
b. Using a simulation approach, approximate the chance the student will pass the exam. (*Hint:* Generate at least 10,000 sets of 25 single-digit numbers. Each number represents the answer to one of the questions, with even numbers recorded as a true answer and odd numbers recorded as a false answer. Determine the relative frequency of 17 or more correct answers in the 25 questions.)

Bus. **4.7** The R&D department of a company has developed a new home screening test for diabetes. A demonstration of the type of results that may occur was mandated by upper management. Simulate the probability of obtaining at least 24 positive results and 6 negative results in a set of 30 results. The researchers state that the probability of obtaining a positive result is 80%.

a. Let a two-digit number represent the outcome of running the screening test. Which numbers should represent a positive result?
b. Approximate the probability of obtaining at least 24 positive results and 6 negative results in a set of 30 results by generating 10,000 sets of 30 two-digit numbers.

Gov. **4.8** The state vehicle inspection bureau provided the following information on the percentage of cars that fail an annual vehicle inspection due to having faulty lights: 15% of all cars have one faulty light, 10% have two faulty lights, and 5% have three or more faulty lights.

a. What is the probability that a randomly selected car will have no faulty lights?
b. What is the probability that a randomly selected car will have at most one faulty light?
c. What is the probability that a randomly selected car will fail an inspection due to a faulty light?

Gov. **4.9** The Texas Lottery has a game, Daily 4, in which a player pays $1 to select four single-digit numbers. Each week the Lottery commission places a set of 10 balls numbered 0–9 in each of four containers. After the balls are thoroughly mixed, one ball is selected from each of the four containers. The winner is the player who matches all four numbers.

a. What is the probability of being the winning player if you purchase a single set of four numbers?
b. Which of the probability approaches (subjective, classical, or relative frequency) did you employ in obtaining your answer in part (a)?

4.3 Basic Event Relations and Probability Laws

Basic **4.10** A die is rolled two times. Provide a list of the possible outcomes of the two rolls in this form: the result from the first roll and the result from the second roll.

Basic **4.11** Refer to Exercise 4.10. Assume that the die is a fair die, that is, each of the outcomes has a probability of 1/36. What is the probability of observing

a. Event A: Exactly one dot appears on each of the two upturned faces?
b. Event B: The sum of the dots on the two upturned faces is exactly 4?
c. Event C: The sum of the dots on the two upturned faces is at most 4?

Basic **4.12** Refer to Exercise 4.11.

a. Describe the event that is the complement of event A.
b. Compute the complement of event A.

Basic **4.13** Refer to Exercise 4.11.

a. Are events A and B mutually exclusive?
b. Are events A and C mutually exclusive?
c. Are events B and C mutually exclusive?

Bus. **4.14** A credit union takes a sample of four mortgages each month to survey the homeowners' satisfaction with the credit union's servicing of their mortgage. Each mortgage is classified as a fixed rate (F) or variable rate (V).

a. What are the 16 possible combinations of the four mortgages? *Hint:* One such possibility would be $F_1V_2V_3F_4$.
b. List the combinations in event A: At least three of the mortgages are variable rate.

c. List the combinations in event B: All four mortgages are the same type.
d. List the combinations in event C: The union of events A and B.
e. List the combinations in event D: The intersection of events A and B.

Engin. **4.15** A nuclear power plant has double redundancy on the feedwater pumps used to remove heat from the reactor core. A safely operating plant requires only one of the three pumps to be functional. Define the events A, B, and C as follows:

A: Pump 1 works properly
B: Pump 2 works properly
C: Pump 3 works properly

Describe in words the following events:

a. The intersection of A, B, and C
b. The union of A, B, and C
c. The complement of the intersection of A, B, and C
d. The complement of the union of A, B, and C

4.16 The population distribution in the United States based on race/ethnicity and blood type as reported by the ***American Red Cross*** is given here.

	Blood Type			
Race/Ethnicity	**O**	**A**	**B**	**AB**
White	36%	32.2%	8.8%	3.2%
Black	7%	2.9%	2.5%	.5%
Asian	1.7%	1.2%	1%	.3%
All others	1.5%	.8%	.3%	.1%

a. A volunteer blood donor walks into a Red Cross blood donation center. What is the probability she will be Asian and have Type O blood?
b. What is the probability that a white donor will not have Type A blood?
c. What is the probability that an Asian donor will have either Type A or Type B blood?
d. What is the probability that a donor will have neither Type A nor Type AB blood?

4.17 The makers of the candy M&Ms report that their plain M&Ms are composed of 15% yellow, 10% red, 20% orange, 25% blue, 15% green, and 15% brown. If you randomly select an M&M, what is the probability of the following?

a. It is brown.
b. It is red or green.
c. It is not blue.
d. It is both red and brown.

4.4 Conditional Probability and Independence

Bus. **4.18** Refer to Exercise 4.11. Compute the following probabilities:

a. $P(A|B)$
b. $P(A|C)$
c. $P(B|C)$

Basic **4.19** Refer to Exercise 4.11.

a. Are the events A and B independent? Why or why not?
b. Are the events A and C independent? Why or why not?
c. Are the events B and C independent? Why or why not?

Basic **4.20** Refer to Exercise 4.14.

a. Are the events A and B independent? Justify your answer.
b. Are the events A and C independent? Justify your answer.
c. Are the events A and D independent? Justify your answer.
d. Which pair(s) of the events are mutually exclusive: (A, B), (B, C,), and/or (A, C)? Justify your answer.

4.21 Refer to Exercise 4.16. Let W be the event that the donor is white, B be the event that the donor is black, and A be the event that the donor is Asian. Also, let T_1 be the event that the donor has blood type O, T_2 be the event that the donor has blood type A, T_3 be the event that the donor has blood type B, and T_4 be the event that the donor has blood type AB.

a. Describe in words the event $T_1|W$.
b. Compute the probability of the occurrence of the event $T_1|W$, $P(T_1|W)$.
c. Are the events W and T_1 independent? Justify your answer.
d. Are the events W and T_1 mutually exclusive? Explain your answer.

4.22 Is it possible for events A and B to be both mutually exclusive and independent? Justify your answer.

H.R. **4.23** A survey of 1,000 U.S. government employees who have an advanced college degree produced the following responses to the offering of a promotion to a higher grade position that would involve moving to a new location.

	Married			
Promotion	**Both Spouses Professional**	**One Spouse Professional**	**Unmarried**	**Total**
Rejected	184	56	17	257
Accepted	276	314	153	743
Total	460	370	170	1,000

Use the results of the survey to estimate the following probabilities.

a. What is the probability that a randomly selected government employee having an advanced college degree would accept a promotion?
b. What is the probability that a randomly selected government employee having an advanced college degree would not accept a promotion?
c. What is the probability that a randomly selected government employee having an advanced college degree has a spouse with a professional position?

H.R. **4.24** Refer to Exercise 4.23. Define the following events.

Event A: A randomly selected government employee having an advanced college degree would accept a promotion

Event B: A randomly selected government employee having an advanced college degree has a spouse in a professional career

Event C: A randomly selected government employee having an advanced college degree has a spouse without a professional position

Event D: A randomly selected government employee having an advanced college degree is unmarried

Use the results of the survey in Exercise 4.23 to compute the following probabilities:

a. $P(A)$
b. $P(B)$
c. $P(A|C)$
d. $P(A|D)$

H.R. **4.25** Refer to Exercise 4.23.

a. Are the events A and C independent? Justify your answer.
b. Are the events A and D independent? Justify your answer.
c. Compute $1 - P(A|B)$ and $P(\overline{A}|B)$. Are they equal?
d. Compute $1 - P(A|B)$ and $P(A|\overline{B})$. Are they equal?

H.R. **4.26** A large corporation has spent considerable time developing employee performance rating scales to evaluate an employee's job performance on a regular basis so major adjustments can be made when needed and employees who should be considered for a "fast track" can be isolated. Keys to this latter determination are ratings on the ability of an employee to perform to his or her capabilities and on his or her formal training for the job.

	Formal Training			
Workload Capacity	**None**	**Little**	**Some**	**Extensive**
Low	.01	.02	.02	.04
Medium	.05	.06	.07	.10
High	.10	.15	.16	.22

The probabilities for being placed on a fast track are as indicated for the 12 categories of workload capacity and formal training. The following three events (A, B, and C) are defined:

A: An employee works at the high-capacity level
B: An employee falls into the highest (extensive) formal training category
C: An employee has little or no formal training and works below high capacity

a. Find $P(A)$, $P(B)$, and $P(C)$.
b. Find $P(A|B)$, $P(B|\overline{B})$, and $P(\overline{B}|C)$.
c. Find $P(A \cup B)$, $P(A \cap C)$, and $P(B \cap C)$.

Bus. **4.27** The utility company in a large metropolitan area finds that 70% of its customers pay a given monthly bill in full.

a. Suppose two customers are chosen at random from the list of all customers. What is the probability that both customers will pay their monthly bill in full?
b. What is the probability that at least one of them will pay in full?

4.28 Refer to Exercise 4.27. A more detailed examination of the company records indicates that 95% of the customers who pay one monthly bill in full will also pay the next monthly bill in full; only 10% of those who pay less than the full amount one month will pay in full the next month.

a. Find the probability that a customer selected at random will pay two consecutive months in full.
b. Find the probability that a customer selected at random will pay neither of two consecutive months in full.
c. Find the probability that a customer chosen at random will pay exactly one month in full.

4.5 Bayes' Formula

Bus. **4.29** Of a finance company's loans, 1% are defaulted (not completely repaid). The company routinely runs credit checks on all loan applicants. It finds that 30% of defaulted loans went to poor risks, 40% to fair risks, and 30% to good risks. Of the nondefaulted loans, 10% went to poor risks, 40% to fair risks, and 50% to good risks. Use Bayes' Formula to calculate the probability that a poor-risk loan will be defaulted.

4.30 Refer to Exercise 4.29. Show that the posterior probability of default, given a fair risk, equals the prior probability of default. Explain why this is a reasonable result.

4.31 In Example 4.4, we described a new test for determining defects in circuit boards. Compute the probability that the test correctly identifies the defects D_1, D_2, and D_3; that is, compute $P(D_1|A_1)$, $P(D_2|A_2)$, and $P(D_3|A_3)$.

4.32 In Example 4.4, compute the probability that the test incorrectly identifies the defects D_1, D_2, and D_3; that is, compute $P(D_1|\overline{A}_1)$, $P(D_2|\overline{A}_2)$, and $P(D_3|\overline{A}_3)$.

Bus. **4.33** An underwriter of home insurance policies studies the problem of home fires resulting from wood-burning furnaces. Of all homes having such furnaces, 30% own a type 1 furnace, 25% a type 2 furnace, 15% a type 3, and 30% other types. Over 3 years, 5% of type 1 furnaces, 3% of type 2, 2% of type 3, and 4% of other types have resulted in fires. If a fire occurs in a particular home, what is the probability that a type 1 furnace is in the home?

Med. **4.34** In a January 15, 1998, article, the ***New England Journal of Medicine (338:141–146)*** reported on the utility of using computerized tomography (CT) as a diagnostic test for patients with clinically suspected appendicitis. In at least 20% of patients with appendicitis, the correct diagnosis was not made. On the other hand, the appendix was normal in 15% to 40% of patients who under-

went emergency appendectomy. A study was designed to determine the prospective effectiveness of using CT as a diagnostic test to improve the treatment of these patients. The study examined 100 consecutive patients suspected of having acute appendicitis who presented to the emergency department or were referred there from a physician's office. The 100 patients underwent a CT scan, and the surgeon made an assessment of the presence of appendicitis for each of the patients. The final clinical outcomes were determined at surgery and by pathological examination of the appendix after appendectomy or by clinical follow-up at least 2 months after CT scanning.

	Presence of Appendicitis	
Radiologic Determination	**Confirmed (C)**	**Ruled Out (RO)**
Definitely appendicitis (DA)	50	1
Equivocally appendicitis (EA)	2	2
Definitely not appendicitis (DNA)	1	44

The 1996 rate of occurrence of appendicitis was approximately $P(C) = .00108$.

a. Find the sensitivity and specificity of the radiological determination of appendicitis.

b. Find the probability that a patient truly had appendicitis given that the radiological determination was definitely appendicitis (DA).

c. Find the probability that a patient truly did not have appendicitis given that the radiological determination was definitely appendicitis (DA).

d. Find the probability that a patient truly did not have appendicitis given that the radiological determination was definitely not appendicitis (DNA).

Med. **4.35** Conditional probabilities can be useful in diagnosing disease. Suppose that three different, closely related diseases (A_1, A_2, and A_3) occur in 25%, 15%, and 12% of the population. In addition, suppose that any one of three mutually exclusive symptom states (B_1, B_2, and B_3) may be associated with each of these diseases. Experience shows that the likelihood $P(B_j|A_i)$ of having a given symptom state when the disease is present is as shown in the following table. Find the probability of disease A_2 given symptoms B_1, B_2, B_3, and B_4, respectively.

	Disease State A_i		
Symptom State B_j	A_1	A_2	A_3
B_1	.08	.17	.10
B_2	.18	.12	.14
B_3	.06	.07	.08
B_4 (no symptoms)	.68	.64	.68

4.6 Variables: Discrete and Continuous

Basic **4.36** Classify each of the following random variables as either continuous or discrete:

a. The survival time of a cancer patient after receiving a new treatment for cancer

b. The number of ticks found on a cow entering an inspection station

c. The average rainfall during August in College Station, Texas

d. The daily dose level of medication prescribed to a patient having an iron deficiency

e. The number of touchdowns thrown during an NFL game

f. The number of monthly shutdowns of the sewage treatment plant in a large midwestern city

Basic **4.37** The U.S. Consumer Product Safey Commission investigates bicycle helmet hazards. The inspectors studied incidents in which deaths resulted from improper uses of helmets. The inspectors recorded the incidents in which children were strangled by the straps on the helmet. Is the number of deaths by helmet strangulation during a randomly selected month a discrete or continuous random variable. Explain your answer.

Basic **4.38** Texting while driving is a very dangerous practice. An electronic monitoring device is installed on rental cars at a randomly selected rental franchise.

a. Is the number of times a randomly selected driver sends a text message during the first hour after leaving the rental company's parking lot a discrete or continuous random variable?

b. Is the length of time the driver spends typing a text message while driving a discrete or continuous random variable?

c. Is the brand of cell phone from which the text message is sent a discrete or continuous random variable?

Basic **4.39** A car dealership uses a questionnaire to evaluate customer interactions with the dealership's salespersons. One of the items on the questionnaire was "Overall, the interaction with the salesperson was positive." The possible responses are *Strongly agree, Agree, No opionion, Disagree,* and *Strongly disagree*.

a. Is the number of customers responding *Strongly agree* a continuous or discrete random variable?

b. Is the proportion of customers responding *Strongly agree* a continuous or discrete random variable?

4.7 Probability Distributions for Discrete Random Variables

Gov. **4.40** The numbers of cars failing an emissions test on randomly selected days at a state inspection station are given in the following table.

y	0	1	2	3	4	5	6	7	8	9	10
$P(y)$	.100	.130	.250	.160	.095	.075	.063	.047	.041	.024	.015

a. Construct a graph of $P(y)$.

b. Compute $P(y \leq 2)$.

c. Compute $P(y > 7)$.

d. Compute $P(2 < y \leq 7)$.

Bus. **4.41** A traditional call center has a simple mission: Agents have to answer customer calls fast and end them as quickly as possible to move on to the next call. The quality of service rendered by the call center was evaluated by recording the number of times a customer called the center back within a week of his or her initial call to the center.

y = number of recalls	0	1	2	3	4	5	6
$P(y)$	.151	.232	.354	.161	.067	.021	.014

a. What is the probability that a customer will recall the center more than three times?

b. What is the probability that a customer will recall the center at least two times but less than five times?

c. Suppose a call center must notify a supervisor if a customer recalls the center more than four times within a week of his or her initial call. What proportion of customers who contact the call center will require a supervisor to be contacted?

4.8 Two Discrete Random Variables: The Binomial and the Poisson

Bio. **4.42** A biologist randomly selects 10 portions of water, each equal to .1 cm^3 in volume, from the local reservoir and counts the number of bacteria present in each portion. The biologist then totals the number of bacteria for the 10 portions to obtain an estimate of the number of bacteria per cubic centimeter present in the reservoir water. Is this a binomial experiment?

Pol. Sci. **4.43** Examine the accompanying newspaper clipping. Does this sampling appear to satisfy the characteristics of a binomial experiment?

Poll Finds Opposition to Phone Taps

New York—People surveyed in a recent poll indicated they are 81% to 13% against having their phones tapped without a court order.

The people in the survey, by 68% to 27%, were opposed to letting the government use a wiretap on citizens suspected of crimes, except with a court order.

The survey was conducted for 1,495 households and also found the following results:

—The people surveyed are 80% to 12% against the use of any kind of electronic spying device without a court order.

—Citizens are 77% to 14% against allowing the government to open their mail without court orders.

—They oppose, by 80% to 12%, letting the telephone company disclose records of long-distance phone calls, except by court order.

For each of the questions, a few of those in the survey had no responses.

Env. **4.44** A survey is conducted to estimate the percentage of pine trees in a forest that are infected by the pine shoot moth. A grid is placed over a map of the forest, dividing the area into 25-foot by 25-foot square sections. One hundred of the squares are randomly selected, and the number of infected trees is recorded for each square. Is this a binomial experiment?

Gov. **4.45** In an attempt to decrease drunk driving, police set up vehicle checkpoints during the July 4 evening. The police randomly select vehicles to be stopped for "informational" checks. On a particular roadway, assume that 20% of all drivers have a blood alcohol level above the legal limit. For a random sample of 15 vehicles, compute the following probabilities:

a. All 15 drivers will have a blood alcohol level exceeding the legal limit.
b. Exactly 6 of the 15 drivers will exceed the legal limit.
c. Of the 15 drivers, 6 or more will exceed the legal limit.
d. All 15 drivers will have a blood alcohol level within the legal limit.

Bus. **4.46** The quality control department examines all the products returned to a store by customers. An examination of the returned products yields the following assessment: 5% are defective and not repairable, 45% are defective but repairable, 35% have small surface scratches but are functioning properly, and 15% have no problems. Compute the following probabilities for a random sample of 20 returned products:

a. All of the 20 returned products have some type of problem.
b. Exactly 6 of the 20 returned products are defective and not repairable.
c. Of the 20 returned products, 6 or more are defective and not functioning properly.
d. None of the 20 returned products has any sort of defect.

Med. **4.47** Knee replacements have emerged as a mainstream surgery. According to the ***Knee Replacement Statistics Agency of Research and Quality (AHRQ)***, over 600,000 procedures were performed in 2009, and the number is expected to grow into the millions by the year 2030. According to the ***American Academy of Orthopedic Surgeons (AAOS)***, serious complications occur in less than 2% of cases. If AAOS is correct that only 2% of knee replacement patients have serious complications, would the next 10 patients at a major teaching hospital receiving a knee replacement constitute a binomial experiment with $n = 10$ and $\pi = .02$? Justify your answer.

Bus. **4.48** The CFO of a hospital is concerned about the risk of patients contracting an infection after a one-week or longer stay in the hospital. A long-term study estimates that the chance of contracting an infection after a one-week or longer stay in a hospital is 10%. A random sample of 50 patients who have been in the hospital at least 1 week is selected.

a. If the 10% infection rate is correct, what is the probability that at least 5 patients out of the 50 will have an infection?
b. What assumptions are you making in computing the probability in part (a)?

Basic **4.49** Suppose the random variable y has a Poisson distribution. Compute the following probabilities:

a. $P(y = 4)$ given $\mu = 2$
b. $P(y = 4)$ given $\mu = 3.5$

c. $P(y > 4)$ given $\mu = 2$
d. $P(1 \leq y < 4)$ given $\mu = 2$

Bus. **4.50** Customers arrive at a grocery store checkout at a rate of six per 30 minutes during the hours of 5 P.M. and 7 P.M. during the workweek. Let C be the number of customers arriving at the checkout during any 30-minute period of time. The management of the store wants to determine the frequency of the following events. Compute the probabilities of these events:

a. No customers arrive.
b. More than six customers arrive.
c. At most three customers arrive.

Bus. **4.51** A firm is considering using the Internet to supplement its traditional sales methods. Using data from an industry association, the firm estimates that 1 of every 1,000 Internet hits results in a sale. Suppose the firm has 2,500 hits per day.

a. What is the probability that the firm will have more than five sales in a randomly selected day?
b. What conditions must be satisfied in order for you to make the calculation in part (a)?
c. Use the Poisson approximation to compute the probability that the firm will have more than five sales in a randomly selected day.
d. Is the Poisson approximation accurate?

4.52 A certain birth defect occurs in 1 of every 10,000 births. In the next 5,000 births at a major hospital, what is the probability that at least 1 baby will have the defect? What assumptions are required to calculate this probability?

4.10 A Continuous Probability Distribution: The Normal Distribution

Basic **4.53** Find the area under the standard normal curve between these values:

a. $z = 0$ and $z = 1.3$
b. $z = 0$ and $z = 2.7$

Basic **4.54** Find the area under the standard normal curve between these values:

a. $z = .5$ and $z = 1.3$
b. $z = -1.3$ and $z = 0$

Basic **4.55** Find the area under the standard normal curve between these values:

a. $z = -2.5$ and $z = -1.2$
b. $z = -1.3$ and $z = -.7$

Basic **4.56** Find the area under the standard normal curve between these values:

a. $z = -1.5$ and $z = 0.2$
b. $z = -1.2$ and $z = 0.7$

In Exercises 4.57 through 4.63, let z be a random variable with a standard normal distribution.

Basic **4.57** Find the probability that z is less than 1.23.

Basic **4.58** Find the probability that z is greater than 0.35.

Basic **4.59** Find the value of z, denoted z_0, such that $P(z < z_0) = .5$.

Basic **4.60** Find the value of z, denoted z_0, such that $P(z > z_0) = .025$.

Basic **4.61** Find the value of z, denoted z_0, such that $P(z > z_0) = .0091$.

Basic **4.62** Find the value of z, denoted z_0, such that $P(-z_0 < z \leq z_0) = .975$.

Basic **4.63** Find the value of z, denoted z_0, such that $P(-z_0 < z \leq z_0) = .90$.

Basic **4.64** Let y be a random variable having a normal distribution with a mean equal to 50 and a standard deviation equal to 8. Find the following probabilities:

a. $P(y > 50)$
b. $P(y > 53)$
c. $P(y < 58)$
d. $P(38 < y < 62)$
e. $P(38 \leq y \leq 62)$

Basic **4.65** Let y be a random variable having a normal distribution with a mean equal to 250 and a standard deviation equal to 50. Find the following probabilities:

a. $P(y > 250)$
b. $P(y > 150)$
c. $P(150 < y < 350)$
d. Find k such that $P(250 - k < y < 250 + k) = .60$

Basic **4.66** Suppose that y is a random variable having a normal distribution with a mean equal to 250 and a standard deviation equal to 10.

a. Show that the event $y < 260$ has the same probability as $z < 1$.
b. Convert the event $y > 230$ to the z-score equivalent.
c. Find $P(y < 260)$ and $P(y > 230)$.
d. Find $P(y > 265)$, $P(y < 242)$, and $P(242 < y < 265)$.

Basic **4.67** Suppose that z is a random variable having a standard normal distribution.

a. Find a value z_0, such that $P(z > z_0) = .01$.
b. Find a value z_0, such that $P(z < z_0) = .025$.
c. Find a value z_0, such that $P(-z_0 < z < z_0) = .95$.

Basic **4.68** Let y be a random variable having a normal distribution with mean equal to 250 and standard deviation equal to 50.

a. Find a value y_0, such that $P(y > y_0) = .01$.
b. Find a value y_0, such that $P(y < y_0) = .025$.
c. Find two values y_1 and y_2, such that $(y_1 + y_2)/2 = 250$ and $P(y_1 < y < y_2) = .95$.

Gov. **4.69** Records maintained by the office of budget in a particular state indicate that the amount of time elapsed between the submission of travel vouchers and the final reimbursement of funds has approximately a normal distribution with a mean of 36 days and a standard deviation of 3 days.

a. What is the probability that the elapsed time between submission and reimbursement will exceed 30 days?
b. If you had a travel voucher submitted more than 55 days ago, what might you conclude?

Edu. **4.70** The College Boards, which are administered each year to many thousands of high school students, are scored so as to yield a mean of 513 and a standard deviation of 130. These scores are close to being normally distributed. What percentage of the scores can be expected to satisfy each of the following conditions?

a. Greater than 600
b. Greater than 700
c. Less than 450
d. Between 450 and 600

Bus. **4.71** Monthly sales figures for a particular food industry tend to be normally distributed with a mean of 155 (thousand dollars) and a standard deviation of 45 (thousand dollars). Compute the following probabilities:

a. $P(y < 200)$
b. $P(y > 100)$
c. $P(100 < y < 200)$

4.72 Refer to Exercise 4.70. An honor society wishes to invite those scoring in the top 5% on the College Boards to join their society.

a. What score is required to be invited to join the society?
b. What score separates the top 75% of the population from the bottom 25%? What do we call this value?

4.11 Random Sampling

Soc. **4.73** City officials want to sample the opinions of the homeowners in a community regarding the desirability of increasing local taxes to improve the quality of the public schools. If a random number table is used to identify the homes to be sampled and a home is discarded if the homeowner is not home when visited by the interviewer, is it likely this process will approximate random sampling? Explain.

Pol. Sci. **4.74** A local TV network wants to run an informal survey of individuals who exit from a local voting station to ascertain early results on a proposal to raise funds to move the city-owned historical museum to a new location. How might the network sample voters to approximate random sampling?

Psy. **4.75** A psychologist is interested in studying women who are in the process of obtaining a divorce to determine whether the women experienced significant attitudinal changes after the divorce has been finalized. Existing records from the geographic area in question show that 798 couples have recently filed for divorce. Assume that a sample of 25 women is needed for the study, and use Table 12 in the Appendix to determine which women should be asked to participate in the study. (*Hint:* Begin in column 2, row 1, and proceed down.)

Pol. Sci. **4.76** Suppose you have been asked to run a public opinion poll related to an upcoming election. There are 230 precincts in the city, and you need to randomly select 50 registered voters from each precinct. Suppose that each precinct has 1,000 registered voters and it is possible to obtain a list of these persons. You assign the numbers 1 to 1,000 to the 1,000 people on each list, with 1 to the first person on the list and 1,000 to the last person. You need to next obtain a random sample of 50 numbers from the numbers 1 to 1,000. The names on the sampling frame corresponding to these 50 numbers will be the 50 persons selected for the poll. Note that you would need to obtain a new random sample for each of the 230 precincts.

- **a.** Using either a random number table or a computer program, generate a random sample of 50 numbers from the numbers 1 to 1,000.
- **b.** Give several reasons why you need to generate a different set of random numbers for each of the precincts. Why not use the same set of 50 numbers for all 230 precincts?

4.12 Sampling Distributions

4.77 A random sample of 16 measurements is drawn from a population with a mean of 60 and a standard deviation of 5. Describe the sampling distribution of $\bar{y}$, the sample mean. Within what interval would you expect $\bar{y}$ to lie approximately 95% of the time?

4.78 Refer to Exercise 4.77. Describe the sampling distribution for the sample sum $\sum y_i$. Is it unlikely (improbable) that $\sum y_i$ would be more than 70 units away from 960? Explain.

Psy. **4.79** Psychomotor retardation scores for a particular group of manic-depressive patients have approximately a normal distribution with a mean of 930 and a standard deviation of 130. A random sample of 20 patients from the group was selected, and their mean psychomotor retardation score was obtained.

- **a.** What is the probability that their mean score was between 900 and 960?
- **b.** What is the probability that their mean score was greater than 960?
- **c.** What is the 90th percentile of their mean scores?

Soc. **4.80** Federal resources have been tentatively approved for the construction of an outpatient clinic. In order to design a facility that will handle patient load requirements and stay within a limited budget, the designers studied patient demand. From studying a similar facility in the area, they found that the distribution of the number of patients requiring hospitalization during a week could be approximated by a normal distribution with a mean of 125 and a standard deviation of 32.

- **a.** Use the Empirical Rule to describe the distribution of y, the number of patients requesting service in a week.
- **b.** If the facility was built with a 160-patient capacity, what fraction of the weeks might the clinic be unable to handle the demand?

4.81 Refer to Exercise 4.80. What size facility should be built so the probability of the patient load's exceeding the clinic capacity is .10? .30?

Soc. **4.82** Based on the 1990 census, the number of hours per day adults spend watching television is approximately normally distributed with a mean of 5 hours and a standard deviation of 1.3 hours.

- **a.** What proportion of the population spends more than 7 hours per day watching television?
- **b.** In a 1998 study of television viewing, a random sample of 500 adults reported that the average number of hours spent viewing television was greater than 5.5 hours

per day. Do the results of this survey appear to be consistent with the 1990 census? (*Hint:* If the census results are still correct, what is the probability that the average viewing time would exceed 5.5 hours?)

Env. **4.83** The level of a particular pollutant, nitrogen oxide, in the exhaust of a hypothetical model of car, the Polluter, when driven in city traffic has approximately a normal distribution with a mean level of 2.1 grams per mile (g/m) and a standard deviation of 0.3 g/m.

- **a.** If the EPA mandates that a nitrogen oxide level of 2.7 g/m cannot be exceeded, what proportion of Polluters would be in violation of the mandate?
- **b.** At most, 25% of Polluters exceed what nitrogen oxide level value (that is, find the 75th percentile)?
- **c.** The company producing the Polluter must reduce the nitrogen oxide level so that at most 5% of its cars exceed the EPA level of 2.7 g/m. If the standard deviation remains 0.3 g/m, to what value must the mean level be reduced so that at most 5% of Polluters would exceed 2.7 g/m?

4.84 Refer to Exercise 4.83. A company has a fleet of 150 Polluters used by its sales staff. Describe the distribution of the total amount, in g/m, of nitrogen oxide produced in the exhaust of this fleet. What are the mean and standard deviation of the total amount, in g/m, of nitrogen oxide in the exhaust for the fleet? (*Hint:* The total amount of nitrogen oxide can be represented as $\sum_{i=1}^{150} W_i$, where W_i is the amount of nitrogen oxide in the exhaust of the ith car. Thus, the Central Limit Theorem for sums is applicable.)

Soc. **4.85** The baggage limit for an airplane is set at 100 pounds per passenger. Thus, for an airplane with 200 passenger seats, there would be a limit of 20,000 pounds. The weight of the baggage of an individual passenger is a random variable with a mean of 95 pounds and a standard deviation of 35 pounds. If all 200 seats are sold for a particular flight, what is the probability that the total weight of the passengers' baggage will exceed the 20,000-pound limit?

Med. **4.86** A patient visits her doctor with concerns about her blood pressure. If the systolic blood pressure exceeds 150, the patient is considered to have high blood pressure, and medication may be prescribed. The problem is that there is a considerable variation in a patient's systolic blood pressure readings during a given day.

- **a.** If a patient's systolic readings during a given day have a normal distribution with a mean of 160 mm mercury and a standard deviation of 20 mm, what is the probability that a single measurement will fail to detect that the patient has high blood pressure?
- **b.** If five measurements are taken at various times during the day, what is the probability that the average blood pressure reading will be less than 150 and hence fail to indicate that the patient has a high blood pressure problem?
- **c.** How many measurements would be required so that the probability of failing to detect that the patient has high blood pressure is at most 1%.

4.13 Normal Approximation to the Binomial

Bus. **4.87** Critical key-entry errors in the data processing operation of a large district bank occur approximately .1% of the time. If a random sample of 10,000 entries is examined, determine the following:

- **a.** The expected number of errors
- **b.** The probability of observing fewer than four errors
- **c.** The probability of observing more than two errors

4.88 Use the binomial distribution with $n = 20$ and $\pi = .5$ to compare the accuracy of the normal approximation to the binomial.

- **a.** Compute the exact probabilities and corresponding normal approximations for $y < 5$.
- **b.** The normal approximation can be improved slightly by taking $P(y \leq 4.5)$. Why should this help? Compare your results.
- **c.** Compute the exact probabilities and corresponding normal approximations with the continuity correction for $P(8 < y < 14)$.

4.89 Let y be a binomial random variable with $n = 10$ and $\pi = .5$.

a. Calculate $P(4 \leq y \leq 6)$.

b. Use a normal approximation without the continuity correction to calculate the same probability. Compare your results. How well did the normal approximation work?

4.90 Refer to Exercise 4.89. Use the continuity correction to compute the probability $P(4 \leq y \leq 6)$. Does the continuity correction help?

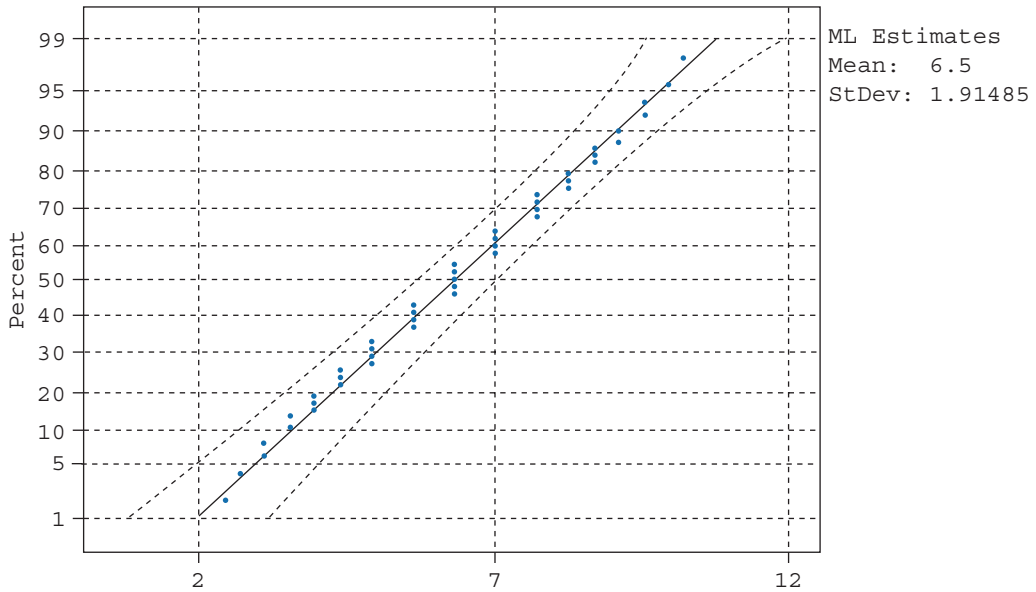
Bus. **4.91** A marketing research firm advises a new client that approximately 15% of all persons sent a sweepstakes offer will return the mailing. Suppose the client sends out 10,000 sweepstakes offers.

a. What is the probability that fewer than 1,430 of the mailings will be returned?

b. What is the probability that more than 1,600 of the mailings will be returned?

4.14 Evaluating Whether or Not a Population Distribution Is Normal

4.92 In Figure 4.19, we visually inspected the relative frequency histogram for sample means based on two measurements and noted its bell shape. Another way to determine whether a set of measurements is bell-shaped (normal) is to construct a **normal probability plot** of the sample data. If the plotted points are nearly a straight line, we say the measurements were selected from a normal population. A normal probability plot was obtained using Minitab software. If the plotted points fall within the curved dotted lines, we consider the data to be a random sample from a normal distribution.



a. Do the 45 data values appear to be a random sample from a normal distribution?

b. Using the values of $\bar{y}$ in Table 4.9, compute the correlation coefficient and p-value for the normal quantile plot to assess whether the data appear to be sampled from a normal distribution.

c. Do the results in part (b) confirm your conclusion from part (a)?

4.93 Suppose a population consists of the 10 measurements (2, 3, 6, 8, 9, 12, 25, 29, 39, 50). Generate the 45 possible values for the sample mean based on a sample of $n = 2$ observations per sample.

a. Use the 45 sample means to determine whether the sampling distribution of the sample mean is approximately normally distributed by constructing a boxplot, relative frequency histogram, and normal quantile plot of the 45 sample means.

b. Compute the correlation coefficient and p-value to assess whether the 45 means appear to be sampled from a normal distribution.
c. Do the results in part (b) confirm your conclusion from part (a)?

4.94 The fracture toughness in concrete specimens is a measure of how likely it is that blocks used in new home construction may fail. A construction investigator obtains a random sample of 15 concrete blocks and determines the following toughness values:

.47, .58, .67, .70, .77, .79, .81, .82, .84, .86, .91, .95, .98, 1.01, 1.04

a. Use a normal quantile plot to assess whether the data appear to fit a normal distribution.
b. Compute the correlation coefficient and p-value for the normal quantile plot. Comment on the degree of fit of the data to a normal distribution.

Supplementary Exercises

Bus. **4.95** One way to audit expense accounts for a large consulting firm is to sample all reports dated the last day of each month. Comment on whether such a sample constitutes a random sample.

Engin. **4.96** The breaking strengths for 1-foot-square samples of a particular synthetic fabric are approximately normally distributed with a mean of 2,250 pounds per square inch (psi) and a standard deviation of 10.2 psi. Find the probability of selecting a 1-foot-square sample of material at random that on testing would have a breaking strength in excess of 2,265 psi.

4.97 Refer to Exercise 4.96. Suppose that a new synthetic fabric has been developed that may have a different mean breaking strength. A random sample of 15 1-foot sections is obtained, and each section is tested for breaking strength. If we assume that the population standard deviation for the new fabric is identical to that for the old fabric, describe the sampling distribution for $\bar{y}$ based on random samples of 15 1-foot sections of new fabric.

4.98 Refer to Exercise 4.97. Suppose that the mean breaking strength for the sample of 15 1-foot sections of the new synthetic fabric is 2,268 psi. What is the probability of observing a value of $\bar{y}$ equal to or greater than 2,268, assuming that the mean breaking strength for the new fabric is 2,250, the same as that for the old?

4.99 Based on your answer in Exercise 4.98, do you believe the new fabric has the same mean breaking strength as the old? (Assume $\sigma = 10.2$.)

Gov. **4.100** Suppose that you are a regional director of an IRS office and that you are charged with sampling 1% of the returns with gross income levels above $15,000. How might you go about this? Would you use random sampling? How?

Med. **4.101** Experts consider high serum cholesterol levels to be associated with an increased incidence of coronary heart disease. Suppose that the natural logarithm of cholesterol levels for males in a given age bracket is normally distributed with a mean of 5.35 and a standard deviation of .12.

a. What percentage of the males in this age bracket could be expected to have a serum cholesterol level greater than 250 mg/ml, the upper limit of the clinical normal range?
b. What percentage of the males could be expected to have serum cholesterol levels within the clinical normal range of 150–250 mg/ml?
c. What percentage of the adult males in this age bracket could be expected to have a very risky cholesterol level—that is, above 300 mg/ml?

Bus. **4.102** Marketing analysts have determined that a particular advertising campaign should make at least 20% of the adult population aware of the advertised product. After a recent campaign, 60 of 400 adults sampled indicated that they had seen the ad and were aware of the new product.

a. Find the approximate probability of observing $y \leq 60$ given that 20% of the population is aware of the product through the campaign.
b. Based on your answer to part (a), does it appear the ad was successful? Explain.

Med. **4.103** One or more specific, minor birth defects occur with probability .0001 (that is, 1 in 10,000 births). If 20,000 babies are born in a given geographic area in a given year, can we calculate the probability of observing at least one of the minor defects using the binomial or normal approximation to the binomial? Explain.

Basic **4.104** The sample mean to be calculated from a random sample of size $n = 4$ from a population that consists of eight measurements (2, 6, 9, 12, 25, 29, 39, 50). Find the sampling distribution of $\bar{y}$. (*Hint:* There are 70 samples of size 4 when sampling from a population of eight measurements.)

Basic **4.105** Plot the sampling distribution of $\bar{y}$ from Exercise 4.104.

a. Does the sampling distribution appear to be approximately normal?

b. Verify that the mean of the sampling distribution of $\bar{y}$ equals the mean of the eight population values.

Basic **4.106** Refer to Exercise 4.104. Use the same population to find the sampling distribution for the sample median based on samples of size $n = 4$.

Basic **4.107** Refer to Exercise 4.106. Plot the sampling distribution of the sample median of Exercise 4.106.

a. Does the sampling distribution appear to be approximately normal?

b. Compute the mean of the sampling distribution of the sample median, and compare this value to the population median.

Basic **4.108** Random samples of size 5, 20, and 80 are drawn from a population with a mean of $\mu = 100$ and a standard deviation of $\sigma = 15$.

a. Give the mean of the sampling distribution of $\bar{y}$ for each of the three sample sizes.

b. Give the standard deviation of the sampling distribution of $\bar{y}$ for each of the three sample sizes.

c. Based on the results obtained in parts (a) and (b), what do you conclude about the accuracy of using the sample mean $\bar{y}$ as an estimate of population mean μ?

Basic **4.109** Refer to Exercise 4.108. To evaluate how accurately the sample mean $\bar{y}$ estimates the population mean μ, we need to know the chance of obtaining a value of $\bar{y}$ that is far from μ. Suppose it is important that the sample mean $\bar{y}$ is within five units of the population mean μ. Find the following probabilities for each of the three sample sizes, and comment on the accuracy of using $\bar{y}$ to estimate μ.

a. $P(\bar{y} \geq 105)$

b. $P(\bar{y} \leq 95)$

c. $P(95 \leq \bar{y} \leq 105)$

Geol. **4.110** Suppose the probability that a major earthquake occurs on a given day in Fresno, California, is 1 in 10,000.

a. In the next 1,000 days, what is the expected number of major earthquakes in Fresno?

b. If the occurrence of major earthquakes can be modeled by the Poisson distribution, calculate the probability that there will be at least one major earthquake in Fresno during the next 1,000 days.

Bio. **4.111** A wildlife biologist is studying turtles that have been exposed to oil spills in the Gulf of Mexico. Previous studies have determined that a particular blood disorder occurs in turtles exposed for a length of time to oil at a rate of 1 in every 8 exposed turtles. The biologist examines 12 turtles exposed for a considerable period of time to oil. If the rate of occurrence of the blood disorder has not changed, what is the probability of each of the following events?

She finds the disorder in

a. None of the 12 turtles.

b. At least 2 of the 12 turtles.

c. No more than 4 turtles.

Bus. **4.112** Airlines overbook (sell more tickets than there are seats) flights, based on past records that indicate that approximately 5% of all passengers fail to arrive on time for their flight. Suppose a plane will hold 250 passengers, but the airline books 260 seats. What is the probability that at least 1 passenger will be bumped from the flight?

Geol. **4.113** For the last 300 years, extensive records have been kept on volcanic activity in Japan. In 2002, there were five eruptions or instances of major seismic activity. From historical records, the

mean number of eruptions or instances of major seismic activity is 2.4 per year. A researcher is interested in modeling the number of eruptions or major seismic activities over the 5-year period of 2005–2010.

a. What probability model might be appropriate?
b. What is the expected number of eruptions or instances of major seismic activity during 2005–2010?
c. What is the probability of no eruptions or instances of major seismic activity during 2005–2010?
d. What is the probability of at least two eruptions or instances of major seismic activity during 2005–2010?

Ecol. **4.114** As part of a study to determine factors that may explain differences in animal species relative to their size, the following body masses (in grams) of 50 different bird species were reported in the paper ***"Temperature and the Northern Distributions of Wintering Birds," by Richard Repasky (1991).***

7.7	10.1	21.6	8.6	12.0	11.4	16.6	9.4
11.5	9.0	8.2	20.2	48.5	21.6	26.1	6.2
19.1	21.0	28.1	10.6	31.6	6.7	5.0	68.8
23.9	19.8	20.1	6.0	99.6	19.8	16.5	9.0
448.0	21.3	17.4	36.9	34.0	41.0	15.9	12.5
10.2	31.0	21.5	11.9	32.5	9.8	93.9	10.9
19.6	14.5						

a. Does the distribution of the body masses appear to follow a normal distribution? Provide both a graphical and a quantitative assessment.
b. Repeat part (a), with the outlier 448.0 removed.
c. Determine the sample mean and median with and without the value 448.0 in the data set.
d. Determine the sample standard deviation and MAD with and without the value 448.0 in the data set.

PART 4

Analyzing the Data, Interpreting the Analyses, and Communicating the Results

CHAPTER 5

Inferences About Population Central Values

5.1 Introduction and Abstract of Research Study

Inference—specifically, decision making and prediction—is centuries old and plays a very important role in our lives. Each of us faces daily personal decisions and situations that require predictions concerning the future. The U.S. government is concerned with the balance of trade with countries in Europe and Asia. An investment advisor wants to know whether inflation will be increasing in the next 6 months. A metallurgist would like to use the results of an experiment to determine whether a new lightweight alloy possesses the strength characteristics necessary for use in automobile manufacturing. A veterinarian investigates the effectiveness of a new chemical for treating heartworm in dogs. The inferences that these individuals make should be based on relevant facts, which we call observations, or data.

In many practical situations, the relevant facts are abundant, seemingly inconsistent, and, in many respects, overwhelming. As a result, a careful decision or prediction is often little better than an outright guess. You need only refer to the "Market Views" section of the *Wall Street Journal* or to one of the financial news shows on cable TV to observe the diversity of expert opinion concerning future stock market behavior. Similarly, a visual analysis of data by scientists and

engineers often yields conflicting opinions regarding conclusions to be drawn from an experiment.

Many individuals tend to feel that their own built-in inference-making equipment is quite good. However, experience suggests that most people are incapable of utilizing large amounts of data, mentally weighing each bit of relevant information, and arriving at a good inference. (You may test your own inference-making ability by using the exercises in Chapters 5 through 10. Scan the data and make an inference before you use the appropriate statistical procedure. Then compare the results.) The statistician, rather than relying upon his or her own intuition, uses statistical results to aid in making inferences. Although we touched on some of the notions involved in statistical inference in preceding chapters, we will now collect our ideas in a presentation of some of the basic ideas involved in statistical inference.

The objective of statistics is to make inferences about a population based on information contained in a sample. Populations are characterized by numerical descriptive measures called *parameters*. Typical population parameters are the mean μ, the median M, the standard deviation σ, and a proportion π. Most inferential problems can be formulated as an inference about one or more parameters of a population. For example, a study is conducted by the Wisconsin Education Department to assess the reading ability of children in the primary grades. The population consists of the scores on a standard reading test of all children in the primary grades in Wisconsin. We are interested in estimating the value of the population mean score μ and the proportion π of scores below a standard, which indicates that a student needs remedial assistance.

estimation
hypothesis testing

Methods for making inferences about parameters fall into one of two categories. Either we will **estimate** the value of the population parameter of interest or we will **test a hypothesis** about the value of the parameter. These two methods of statistical inference—estimation and hypothesis testing—involve different procedures, and, more important, they answer two different questions about the parameter. In estimating a population parameter, we are answering the question "What is the value of the population parameter?" In testing a hypothesis, we are seeking an answer to the question "Does the population parameter satisfy a specified condition—for example, '$\mu > 20$' or '$\pi < .3$'?"

Consider a study in which an investigator wishes to examine the effectiveness of a drug product in reducing anxiety levels of anxious patients. The investigator uses a screening procedure to identify a group of anxious patients. After the patients are admitted into the study, each one's anxiety level is measured on a rating scale immediately before he or she receives the first dose of the drug and then at the end of 1 week of drug therapy. These sample data can be used to make inferences about the population from which the sample was drawn either by estimation or by a statistical test:

Estimation:	Information from the sample can be used to estimate the mean decrease in anxiety ratings for the set of all anxious patients who may conceivably be treated with the drug.
Statistical test:	Information from the sample can be used to determine whether the population mean decrease in anxiety ratings is greater than zero.

Notice that the inference related to estimation is aimed at answering the question "What is the mean decrease in anxiety ratings for the population?" In contrast, the statistical test attempts to answer the question "Is the mean drop in anxiety ratings greater than zero?"

Abstract of Research Study: Percentage of Calories from Fat

There has been an increased recognition of the potential relationship between diet and certain diseases. Substantial differences in the rate of incidence of breast cancer across international boundaries and changes in incidence rates as people migrate from low-incidence to high-incidence areas indicates that environmental factors, such as diet, may play a role in the occurrence of certain types of diseases. For example, the percentage of calories from fat in the diet may be related to the incidence of certain types of cancer and heart disease. Recommendations by federal health agencies to reduce fat intake to approximately 30% of total calories are partially based on studies that forecast a reduced incidence of heart disease and breast cancer. The cover and lead article in the August 23, 2004, issue of ***Newsweek*** were titled ***"What You Don't Know About Fat."*** The article details the mechanisms by which fat cells swell to as much as six times their normal size and begin to multiply, from 40 billion in an average adult to 100 billion, when calorie intake greatly exceeds expenditures of calories through exercise. Fat cells require enormous amounts of blood (in comparison to an equal weight of lean muscle), which places a strain on the cardiovascular system. Obesity results in increased wear on the joints, leading to osteoarthritis. Fat cells also secrete estrogen, which has been linked to breast cancer in postmenopausal women. Type 2 (adult-onset) diabetes has as one of its major risk factors obesity. Researchers suspect that the origin of diabetes lies at least partially in the biochemistry of fat. The article states that the evidence that obesity is bad for you is statistical and unassailable. The problem is that some leading companies in the food industry contest some of the claims made linking obesity to health problems based on the fact that it is statistical evidence. Thus, research in laboratories and retrospective studies of people's diet continue in order to provide needed evidence to convince governmental agencies and the public that a major change in people's diet is a necessity.

The assessment and quantification of a person's usual diet is crucial in evaluating the degree of relationship between diet and diseases. This is a very difficult task, but it is important in an effort to monitor dietary behavior among individuals. ***Rosner, Willett, and Spiegelman, in "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Systematic Within-Person Measurement Error" [Statistics in Medicine (1989) 8:1051–1070],*** describe a nurses' health study in which the diet of a large sample of women was examined. Nurses receive information about effects of dietary fat on health in nutrition courses taken as a part of their training. One of the objectives of the study was to determine the percentage of calories from fat in the diet of a population of nurses and compare this value with the recommended value of 30%. This would assist nursing instructors in determining the impact of the material learned in nutritionally related courses on the nurses' personal dietary decisions. There are many dietary assessment methodologies. The most commonly used method in large nutritional epidemiology studies is the food frequency questionnaire (FFQ). This questionnaire uses a carefully designed series of questions to determine the dietary intakes of participants in the study. In the nurses' health study, a sample of nurses completed a single FFQ. These women represented a random sample from a population of nurses. From the information gathered from the questionnaire, the percentage of calories from fat (PCF) was computed. The parameters of interest were the average PCF value, μ for the population of nurses, the standard deviation σ of PCF for the population of nurses, and the proportion π of nurses having PCF greater than 50%, as well as other parameters. The number of subjects needed in the study was determined by specifying the necessary degree of accuracy in the estimation of

the parameters μ, σ, and π. We will discuss in later sections in this chapter several methods for determining the proper sample sizes. For this study, it was decided that a sample of 168 participants would be adequate. The data is given in Section 5.10. The researchers were interested in estimating the parameters associated with PCF along with providing an assessment of how accurately the sample estimators represented the parameters for the whole population. An important question of interest to the researchers was whether the average PCF for the population exceeded the current recommended value of 30%. If the average value is 32% for the sample of nurses, what can we conclude about the average value for the population of nurses? At the end of this chapter, we will provide an answer to this question, along with other results and conclusions reached in this research study.

5.2 Estimation of μ

The first step in statistical inference is point estimation, in which we compute a single value (statistic) from the sample data to estimate a population parameter. Suppose that we are interested in estimating a population mean and that we are willing to assume the underlying population is normal. One natural statistic that could be used to estimate the population mean is the sample mean, but we also could use the median and the trimmed mean. Which sample statistic should we use?

A whole branch of mathematical statistics deals with problems related to developing point estimators (the formulas for calculating specific point estimates from sample data) of parameters from various underlying populations and determining whether a particular point estimator has certain desirable properties. Fortunately, we will not have to derive these point estimators—they'll be given to us for each parameter. When we know which point estimator (formula) to use for a given parameter, we can develop confidence intervals (interval estimates) for these same parameters.

In this section, we deal with point and interval estimation of a population mean μ. Tests of hypotheses about μ are covered in Section 5.4.

For most problems in this text, we will use sample mean $\bar{y}$ as a point estimate of μ; we also will use it to form an interval estimate for the population mean μ. From the Central Limit Theorem for the sample mean (Chapter 4), we know that for a large n, $\bar{y}$ will be approximately normally distributed, with a mean μ and a standard error $\sigma/\sqrt{n}$. Then from our knowledge of the Empirical Rule and areas under a normal curve, we know that the interval $\mu \pm 2\sigma/\sqrt{n}$, or, more precisely, the interval $\mu \pm 1.96\sigma/\sqrt{n}$, includes 95% of the $\bar{y}$s in repeated sampling, as shown in Figure 5.1.

From Figure 5.1, we can observe that the sample mean $\bar{y}$ may not be very close to the population mean μ, the quantity it is supposed to estimate. Thus, when the value of $\bar{y}$ is reported, we should also provide an indication of how accurately $\bar{y}$ estimates μ.

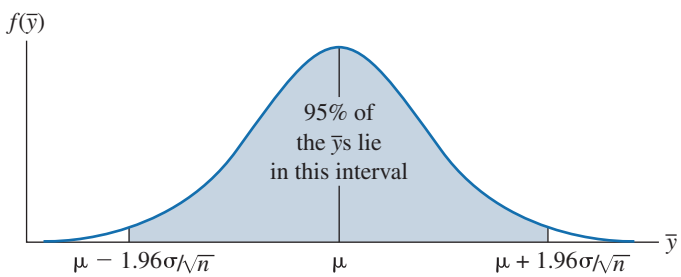
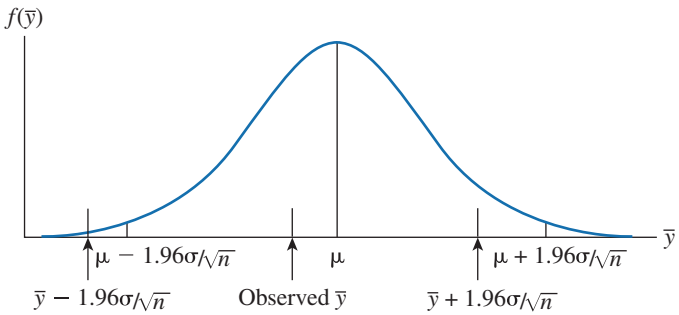


FIGURE 5.1
Sampling distribution for $\bar{y}$

FIGURE 5.2
When the observed value of $\bar{y}$ lies in the interval $\mu \pm 1.96\sigma/\sqrt{n}$, the interval $\bar{y} \pm 1.96\sigma/\sqrt{n}$ contains the parameter μ



We will accomplish this by considering an interval of possible values for μ in place of using just a single value $\bar{y}$. Consider the interval $\bar{y} \pm 1.96\sigma/\sqrt{n}$. Any time $\bar{y}$ falls in the interval $\mu \pm 1.96\sigma/\sqrt{n}$, the interval $\bar{y} \pm 1.96\sigma/\sqrt{n}$ will contain the parameter μ (see Figure 5.2). The probability of $\bar{y}$ falling in the interval $\mu \pm 1.96\sigma/\sqrt{n}$ is .95, so we state that $\bar{y} \pm 1.96\sigma/\sqrt{n}$ is an **interval estimate** of μ with **level of confidence** .95.

interval estimate
level of confidence

confidence coefficient

We evaluate the goodness of an interval estimation procedure by examining the fraction of times in repeated sampling that interval estimates would encompass the parameter to be estimated. This fraction, called the **confidence coefficient,** is .95 when using the formula $\bar{y} \pm 1.96\sigma/\sqrt{n}$; that is, 95% of the time in repeated sampling the intervals calculated using the formula $\bar{y} \pm 1.96\sigma/\sqrt{n}$ will contain the mean μ.

This idea is illustrated in Figure 5.3. Suppose we want to study a commercial process that produces shrimp for sale to restaurants. The shrimp are monitored for size by randomly selecting 40 shrimp from the tanks and measuring their length. We will consider a simulation of the shrimp monitoring. Suppose that the distribution of shrimp length in the tank had a normal distribution with a mean $\mu = 27$ cm and a standard deviation $\sigma = 10$ cm. One hundred samples of size $n = 40$ are drawn from the shrimp population. From each of these samples, we compute the interval estimate $\bar{y} \pm 1.96\sigma/\sqrt{n} = \bar{y} \pm 1.96(10/\sqrt{40})$. (See Table 5.1.) Note that although the intervals vary in location, only 6 of the 100 intervals failed to capture the population mean μ. The fact that six samples produced intervals that did not contain μ is not an indication that the procedure for producing intervals is faulty. Because our level of confidence is 95%, we would expect that, in a large

FIGURE 5.3
Fifty interval estimates of the population mean (27)

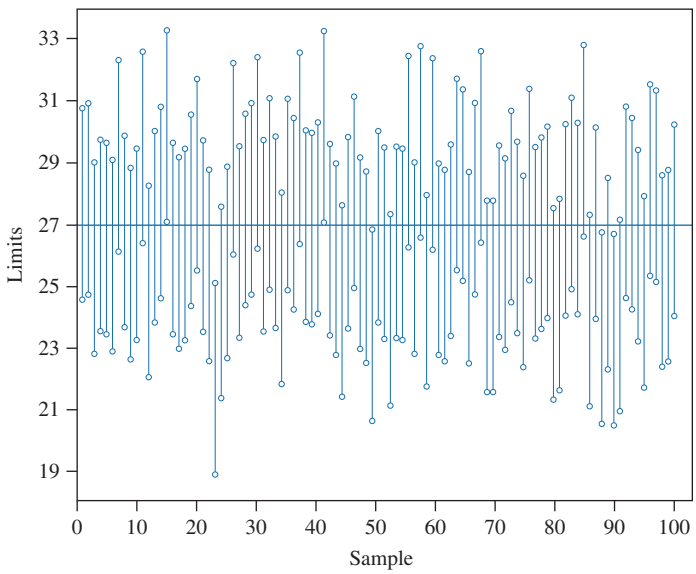


TABLE 5.1 One hundred interval estimates of the population mean (27)

Sample	Sample Mean	Lower Limit	Upper Limit	Interval Contains Population Mean	Sample	Sample Mean	Lower Limit	Upper Limit	Interval Contains Population Mean
1	27.6609	24.5619	30.7599	Yes	51	26.9387	23.8397	30.0377	Yes
2	27.8315	24.7325	30.9305	Yes	52	26.4229	23.3239	29.5219	Yes
3	25.9366	22.8376	29.0356	Yes	53	24.2275	21.1285	27.3265	Yes
4	26.6584	23.5594	29.7574	Yes	54	26.4426	23.3436	29.5416	Yes
5	26.5366	23.4376	29.6356	Yes	55	26.3718	23.2728	29.4708	Yes
6	25.9903	22.8913	29.0893	Yes	56	29.3690	26.2700	32.4680	Yes
7	29.2381	26.1391	32.3371	Yes	57	25.9233	22.8243	29.0223	Yes
8	26.7698	23.6708	29.8688	Yes	58	29.6878	26.5888	32.7868	Yes
9	25.7277	22.6287	28.8267	Yes	59	24.8782	21.7792	27.9772	Yes
10	26.3698	23.2708	29.4688	Yes	60	29.2868	26.1878	32.3858	Yes
11	29.4980	26.3990	32.5970	Yes	61	25.8719	22.7729	28.9709	Yes
12	25.1405	22.0415	28.2395	Yes	62	25.6650	22.5660	28.7640	Yes
13	26.9266	23.8276	30.0256	Yes	63	26.4958	23.3968	29.5948	Yes
14	27.7210	24.6220	30.8200	Yes	64	28.6329	25.5339	31.7319	Yes
15	30.1959	27.0969	33.2949	No	65	28.2699	25.1709	31.3689	Yes
16	26.5623	23.4633	29.6613	Yes	66	25.6491	22.5501	28.7481	Yes
17	26.0859	22.9869	29.1849	Yes	67	27.8394	24.7404	30.9384	Yes
18	26.3585	23.2595	29.4575	Yes	68	29.5261	26.4271	32.6251	Yes
19	27.4504	24.3514	30.5494	Yes	69	24.6784	21.5794	27.7774	Yes
20	28.6304	25.5314	31.7294	Yes	70	24.6646	21.5656	27.7636	Yes
21	26.6415	23.5425	29.7405	Yes	71	26.4696	23.3706	29.5686	Yes
22	25.6783	22.5793	28.7773	Yes	72	26.0308	22.9318	29.1298	Yes
23	22.0290	18.9300	25.1280	No	73	27.5731	24.4741	30.6721	Yes
24	24.4749	21.3759	27.5739	Yes	74	26.5938	23.4948	29.6928	Yes
25	25.7687	22.6697	28.8677	Yes	75	25.4701	22.3711	28.5691	Yes
26	29.1375	26.0385	32.2365	Yes	76	28.3079	25.2089	31.4069	Yes
27	26.4457	23.3467	29.5447	Yes	77	26.4159	23.3169	29.5149	Yes
28	27.4909	24.3919	30.5899	Yes	78	26.7439	23.6449	29.8429	Yes
29	27.8137	24.7147	30.9127	Yes	79	27.0831	23.9841	30.1821	Yes
30	29.3100	26.2110	32.4090	Yes	80	24.4346	21.3356	27.5336	Yes
31	26.6455	23.5465	29.7445	Yes	81	24.7468	21.6478	27.8458	Yes
32	27.9707	24.8717	31.0697	Yes	82	27.1649	24.0659	30.2639	Yes
33	26.7505	23.6515	29.8495	Yes	83	28.0252	24.9262	31.1242	Yes
34	24.9366	21.8376	28.0356	Yes	84	27.1953	24.0963	30.2943	Yes
35	27.9943	24.8953	31.0933	Yes	85	29.7399	26.6409	32.8389	Yes
36	27.3375	24.2385	30.4365	Yes	86	24.2036	21.1046	27.3026	Yes
37	29.4787	26.3797	32.5777	Yes	87	27.0769	23.9779	30.1759	Yes
38	26.9669	23.8679	30.0659	Yes	88	23.6720	20.5730	26.7710	No
39	26.9031	23.8041	30.0021	Yes	89	25.4356	22.3366	28.5346	Yes
40	27.2275	24.1285	30.3265	Yes	90	23.6151	20.5161	26.7141	No
41	30.1865	27.0875	33.2855	No	91	24.0929	20.9939	27.1919	Yes
42	26.4936	23.3946	29.5926	Yes	92	27.7310	24.6320	30.8300	Yes
43	25.8962	22.7972	28.9952	Yes	93	27.3537	24.2547	30.4527	Yes
44	24.5377	21.4387	27.6367	Yes	94	26.3139	23.2149	29.4129	Yes
45	26.1798	23.0808	29.2788	Yes	95	24.8383	21.7393	27.9373	Yes
46	26.7470	23.6480	29.8460	Yes	96	28.4564	25.3574	31.5554	Yes
47	28.0406	24.9416	31.1396	Yes	97	28.2395	25.1405	31.3385	Yes
48	26.0824	22.9834	29.1814	Yes	98	25.5058	22.4068	28.6048	Yes
49	25.6270	22.5280	28.7260	Yes	99	25.6857	22.5867	28.7847	Yes
50	23.7449	20.6459	26.8439	No	100	27.1540	24.0550	30.2530	Yes

collection of 95% confidence intervals, approximately 5% of the intervals would fail to include μ. Thus, in 100 intervals, we would expect 4 to 6 intervals (5% of 100) to not contain μ. It is crucial to understand that even when experiments are properly conducted, a number of the experiments will yield results that in some sense are in error. This occurs when we run only a small number of experiments or select only a small subset of the population. In our example, we randomly selected 40 observations from the population and then constructed a 95% confidence interval for the population mean μ. If this process was repeated a very large number of times—for example, 10,000 times instead of the 100 in our example—the proportion of intervals containing μ would be very nearly 95%.

In most situations when the population mean is unknown, the population standard deviation σ will also be unknown. Hence, it will be necessary to estimate both μ and σ from the data. However, for all practical purposes, if the sample size is relatively large, we can estimate the population standard deviation σ with the sample standard deviation s in the confidence interval formula. Because σ is estimated by the sample standard deviation s, the actual standard error of the mean $\sigma/\sqrt{n}$ is naturally estimated by $s/\sqrt{n}$. This estimation introduces another source of random error (s will vary randomly, from sample to sample, about σ) and, strictly speaking, invalidates the level of confidence for our interval estimate of μ. Fortunately, the formula is still a very good approximation for large sample sizes. When the population has a normal distribution, a better method for constructing the confidence interval will be presented in Section 5.7. Also, based on the results from the Central Limit Theorem, if the population distribution is not too nonnormal and the sample size is relatively large, level of confidence for the interval $\bar{y} \pm 1.96s/\sqrt{n}$ will be approximately the same as if we were sampling from a normal distribution with σ known and using the interval $\bar{y} \pm 1.96\sigma/\sqrt{n}$.

EXAMPLE 5.1

A courier company in New York City claims that its mean delivery time to any place in the city is less than 3 hours. The consumer protection agency decides to conduct a study to see if this claim is true. The agency randomly selects 50 deliveries and determines the mean delivery time to be 2.8 hours with a standard deviation of $s = .6$ hours. The agency wants to estimate the mean delivery time μ using a 95% confidence interval. Obtain this interval and then decide if the courier company's claim appears to be reasonable.

Solution The random sample of $n = 50$ deliveries yields $\bar{y} = 2.8$ and $s = .6$. Because the sample size is relatively large, $n = 50$, the appropriate 95% confidence interval is then computed using the following formula:

$$\bar{y} \pm 1.96\sigma/\sqrt{n}$$

With s used as an estimate of σ, our 95% confidence interval is

$$2.8 \pm 1.96\frac{.6}{\sqrt{50}} \quad \text{or} \quad 2.8 \pm .166$$

The interval from 2.634 to 2.966 forms a 95% confidence interval for the mean delivery time, μ· In other words, we are 95% confident that the average delivery time lies between 2.634 and 2.966 hours. Because the upper value of this interval, 2.966, is less than 3 hours, we can conclude that the data strongly support the courier company's claim. ■

There are many different confidence intervals for μ, depending on the confidence coefficient we choose. For example, the interval $\mu \pm 2.58\sigma/\sqrt{n}$ includes 99% of the values of $\bar{y}$ in repeated sampling, and the interval $\bar{y} \pm 2.58\sigma/\sqrt{n}$ forms a **99% confidence interval** for μ.

99% confidence interval
$(1 - \alpha)$ = confidence coefficient

We can state a general formula for a confidence interval for μ with a **confidence coefficient of $(1 - \alpha)$,** where α (Greek letter alpha) is between 0 and 1. For a specified value of $(1 - \alpha)$, a $100(1 - \alpha)\%$ confidence interval for μ is given by the following formula. Here we assume that σ is known or that the sample size is large enough to replace σ with s.

Confidence Interval for μ, σ Known

$$\bar{y} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

$z_{\alpha/2}$

The quantity $z_{\alpha/2}$ is a value of z having a tail area of $\alpha/2$ to its right. In other words, at a distance of $z_{\alpha/2}$ standard deviations to the right of μ, there is an area of $\alpha/2$ under the normal curve. Values of $z_{\alpha/2}$ can be obtained from Table 1 in the Appendix by looking up the z-value corresponding to an area of $1 - (\alpha/2)$ (see Figure 5.4). Common values of the confidence coefficient $(1 - \alpha)$ and $z_{\alpha/2}$ are given in Table 5.2.

FIGURE 5.4
Interpretation of $z_{\alpha/2}$ in the confidence interval formula

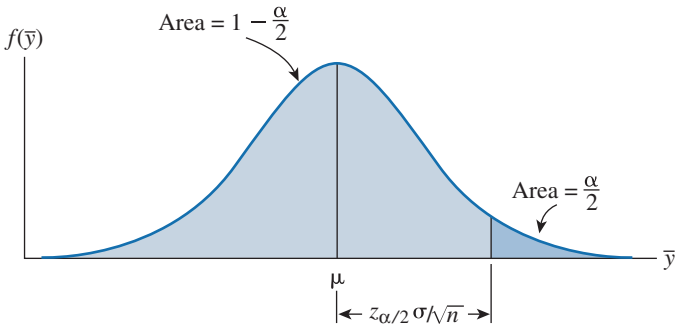


TABLE 5.2
Common values of the confidence coefficient $(1 - \alpha)$ and the corresponding z-value, $z_{\alpha/2}$

Confidence Coefficient $(1 - \alpha)$	Value of $\alpha/2$	Area in Table 1 $1 - \alpha/2$	Corresponding z-Value, $z_{\alpha/2}$
.90	.05	.95	1.645
.95	.025	.975	1.96
.98	.01	.99	2.33
.99	.005	.995	2.58

EXAMPLE 5.2

A forester wishes to estimate the average number of "count trees" (trees larger than a specified size) per acre on a 2,000-acre plantation. She can then use this information to determine the total timber volume for trees in the plantation. A random sample of $n = 50$ 1-acre plots is selected and examined. The average (mean) number of count trees per acre is found to be 27.3, with a standard deviation of 12.1. Use this information to construct a 99% confidence interval for μ, the mean number of count trees per acre for the entire plantation.

Solution We use the general confidence interval with a confidence coefficient equal to .99 and a $z_{\alpha/2}$-value equal to 2.58 (see Table 5.2). Substituting into the formula $\bar{y} \pm 2.58\,\sigma/\sqrt{n}$ and replacing σ with s, we have

$$27.3 \pm 2.58\frac{12.1}{\sqrt{50}}$$

This corresponds to the confidence interval 27.3 ± 4.41—that is, the interval from 22.89 to 31.71. Thus, we are 99% sure that the average number of count trees per acre is between 22.89 and 31.71. ■

Statistical inference-making procedures differ from ordinary procedures in that we not only make an inference but also provide a measure of how good that inference is. For interval estimation, the width of the confidence interval and the confidence coefficient measure the goodness of the inference. For a given value of the confidence coefficient, the smaller the width of the interval, the more precise the inference. The confidence coefficient, on the other hand, is set by the experimenter to express how much confidence he or she has that the interval estimate encompasses the parameter of interest. For a fixed sample size, increasing the level of confidence will result in an interval of greater width. Thus, the experimenter will generally express a desired level of confidence and specify the desired width of the interval. Next, we will discuss a procedure to determine the appropriate sample size to meet these specifications.

5.3 Choosing the Sample Size for Estimating μ

How can we determine the number of observations to include in the sample? The implications of such a question are clear. Data collection costs money. If the sample is too large, time and talent are wasted. Conversely, it is wasteful if the sample is too small because inadequate information has been purchased for the time and effort expended. Also, it may be impossible to increase the sample size at a later time. Hence, the number of observations to be included in the sample will be a compromise between the desired accuracy of the sample statistic as an estimate of the population parameter and the required time and cost to achieve this degree of accuracy.

The researchers in the dietary study described in Section 5.1 had to determine how many nurses to survey for their study to yield viable conclusions. To determine how many nurses must be sampled, we have to determine how accurately the researchers want to estimate the mean percentage of calories from fat (PCF). The researchers specified that they wanted the sample estimator to be within 1.5 of the population mean μ. Then we would want the confidence interval for μ to be $\bar{y} \pm 1.5$. Alternatively, the researchers could specify that the tolerable error in estimation is 3, which would yield the same specification $\bar{y} \pm 1.5$ because the tolerable error is simply the width of the confidence interval.

There are two considerations in determining the appropriate sample size for estimating μ using a confidence interval. First, the tolerable error establishes the desired width of the interval. The second consideration is the level of confidence. In selecting our specifications, we need to consider that if the confidence interval of μ is too wide, then our estimation of μ will be imprecise and not very informative. Similarly, a very low level of confidence (say, 50%) will yield a confidence interval that very likely will be in error—that is, fail to contain μ. However, obtaining a confidence interval having a narrow width and a high level of confidence may require a large value for the sample size and hence be unreasonable in terms of cost and/or time.

What constitutes reasonable certainty? In most situations, the confidence level is set at 95% or 90%, partly because of tradition and partly because these levels represent (to some people) a reasonable level of certainty. The 95% (or 90%) level translates into a long-run chance of 1 in 20 (or 1 in 10) of not covering the population parameter. This seems reasonable and is comprehensible, whereas 1 chance in 1,000 or 1 in 10,000 is too small.

The tolerable error depends heavily on the context of the problem, and only someone who is familiar with the situation can make a reasonable judgment about its magnitude.

When considering a confidence interval for a population mean μ, the plus-or-minus term of the confidence interval is $z_{\sigma/2}\sigma/\sqrt{n}$. Three quantities determine the value of the plus-or-minus term: the desired confidence level (which determines the z-value used), the standard deviation (σ), and the sample size. Usually, a guess must be made about the size of the population standard deviation. An initial sample can be taken to estimate the standard deviation; or the value of the sample standard deviation from a previous study can be used as an estimate of σ. For a given tolerable error, once the confidence level is specified and an estimate of σ supplied, the required sample size can be calculated using the formula shown here.

Suppose we want to estimate μ using a $100(1 - \alpha)\%$ confidence interval having tolerable error W. Our interval will be of the form $\bar{y} \pm E$, where $E = W/2$. Note that W is the width of the confidence interval. To determine the sample size n, we solve the equation

$$E = z_{\sigma/2}\sigma/\sqrt{n}$$

for n. This formula for n is shown here:

Sample Size Required for a $100(1 - \alpha)\%$ Confidence Interval for μ of the Form $\bar{y} \pm E$

$$n = \frac{(z_{\sigma/2})^2\sigma^2}{E^2}$$

Note that determining a sample size to estimate μ requires knowledge of the population standard deviation σ. We can obtain an approximate sample size by estimating σ^2, using one of these two methods:

1. Employ information from a prior experiment to calculate a sample standard deviation s. This value is used to approximate σ.
2. Use information on the range of the observations in the population to obtain an estimate of σ.

We can then substitute the estimated value of σ in the sample-size equation to determine an approximate sample size n.

We illustrate the procedure for choosing a sample size with two examples.

EXAMPLE 5.3

The cost of textbooks relative to other academic expenses has risen greatly over the past few years, and university officials have started to include the average amount expended on textbooks in their estimated yearly expenses for students. In order for these estimates to be useful, they should be within $25 of the mean expenditure for all undergraduate students at the university. How many students should the university sample in order to be 95% confident that its estimated cost of textbooks will satisfy the stated level of accuracy?

Solution From data collected in previous years, the university officials have determined that the annual expenditure for textbooks has a histogram that is normal in shape with costs ranging from \$250 to \$750. An estimate of σ is required to find the sample size. Because the distribution of book expenditures has a normal-like shape, a reasonable estimate of σ would be

$$\hat{\sigma} = \frac{\text{range}}{4} = \frac{750 - 250}{4} = 125$$

The various components in the sample size formula are level of accuracy $= E =$ \$25, $\hat{\sigma} = 125$, and level of confidence $= 95\%$ which implies $z_{\alpha/2} = z_{.05/2} = z_{.025} = 1.96$. Substituting into the sample-size formula, we have

$$n = \frac{(1.96)^2(125)^2}{(25)^2} = 96.04$$

To be on the safe side, we round this number up to the next integer. A sample size of 97 or larger is recommended to obtain an estimate of the mean textbook expenditure that we are 95% confident is within \$25 of the true mean. ■

EXAMPLE 5.4

A federal agency has decided to investigate the advertised weight printed on cartons of a certain brand of cereal. The company in question periodically samples cartons of cereal coming off the production line to check their weight. A summary of 1,500 of the weights made available to the agency indicates a mean weight of 11.80 ounces per carton and a standard deviation of .75 ounce. Use this information to determine the number of cereal cartons the federal agency must examine to estimate the average weight of cartons being produced now, using a 99% confidence interval of width .50.

Solution The federal agency has specified that the width of the confidence interval is to be .50, so $E = .25$. Assuming that the weights made available to the agency by the company are accurate, we can take $\sigma = .75$. The required sample size with $z_{\alpha/2} = 2.58$ is

$$n = \frac{(2.58)^2(.75)^2}{(.25)^2} = 59.91$$

Thus, the federal agency must obtain a random sample of 60 cereal cartons to estimate the mean weight to within $\pm .25$. ■

5.4 A Statistical Test for μ

A second type of inference-making procedure is statistical testing (or hypothesis testing). As with estimation procedures, we will make an inference about a population parameter, but here the inference will be of a different sort. With point and interval estimates, there was no supposition about the actual value of the parameter prior to collecting the data. Using sampled data from the population, we are simply attempting to determine the value of the parameter. In hypothesis testing, there is a preconceived idea about the value of the population parameter. For example, in studying the antipsychotic properties of an experimental compound,

we might ask whether the average shock-avoidance response of rats treated with a specific dose of the compound is greater than 60—that is, $\mu > 60$—the value that has been observed after extensive testing using a suitable standard drug. Thus, there are two theories or hypotheses involved in a statistical study. The first is the hypothesis being proposed by the person conducting the study, called the **research hypothesis**—$\mu > 60$ in our example. The second theory is the negation of this hypothesis, called the **null hypothesis**—$\mu \leq 60$ in our example. The goal of the study is to decide whether the data tend to support the research hypothesis.

research hypothesis

null hypothesis

statistical test

A **statistical test** is based on the concept of proof by contradiction and is composed of the five parts listed here.

1. Research hypothesis (also called the alternative hypothesis), denoted by H_a.
2. Null hypothesis, denoted by H_0.
3. Test statistics, denoted by T.S.
4. Rejection region, denoted by R.R.
5. Check assumptions and draw conclusions.

For example, the Texas A&M agricultural extension service wants to determine whether the mean yield per acre (in bushels) for a particular variety of soybeans has increased during the current year over the mean yield in the previous 2 years when μ was 520 bushels per acre. The first step in setting up a statistical test is determining the proper specification of H_0 and H_a. The following guidelines will be helpful:

1. The statement that μ equals a specific value will always be included in H_0. The particular value specified for μ is called its null value and is denoted μ_0.
2. The statement about μ that the researcher is attempting to support or detect with the data from the study is the research hypothesis, H_a.
3. The negation of H_a is the null hypothesis, H_0.
4. The null hypothesis is presumed correct unless there is overwhelming evidence in the data that the research hypothesis is supported.

In our example, μ_0 is 520. The research statement is that yield in the current year has increased above 520; that is, H_a: $\mu > 520$. (Note that we will include 520 in the null hypothesis.) Thus, the null hypothesis, the negation of H_a, is H_0: $\mu \leq 520$.

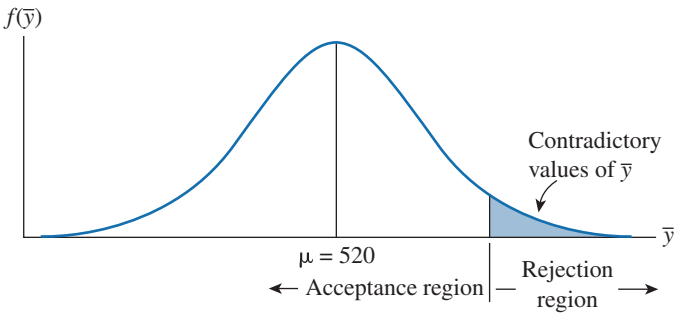
To evaluate the research hypothesis, we take the information in the sample data and attempt to determine whether the data support the research hypothesis or the null hypothesis, but we will give the benefit of the doubt to the null hypothesis.

After stating the null and research hypotheses, we then obtain a random sample of 1-acre yields from farms throughout the state. The decision to state whether or not the data support the research hypothesis is based on a quantity computed from the sample data called the **test statistic.** If the population distribution is determined to be mound-shaped, a logical choice as a test statistic for μ is $\bar{y}$ or some function of $\bar{y}$.

test statistic

If we select $\bar{y}$ as the test statistic, we know that the sampling distribution of $\bar{y}$ is approximately normal with a mean μ and a standard deviation $\sigma/\sqrt{n}$, provided the population distribution is normal or the sample size is fairly large. We are attempting to decide between H_a: $\mu > 520$ and H_0: $\mu \leq 520$. The decision will be to either reject H_0 or fail to reject H_0. In developing our decision rule, we will assume

FIGURE 5.5
Assuming that H_0 is true, contradictory values of $\bar{y}$ are in the upper tail



rejection region

that $\mu = 520$, the null value of μ. We will now determine the values of $\bar{y}$ that define what is called the **rejection region;** we are very unlikely to observe these values if $\mu = 520$ (or if μ is any other value in H_0). The rejection region contains the values of $\bar{y}$ that support the research hypothesis and contradict the null hypothesis; hence, it is the region of values for $\bar{y}$ that reject the null hypothesis. The rejection region will be the values of $\bar{y}$ in the upper tail of the null distribution ($\mu = 520$) of $\bar{y}$. See Figure 5.5.

Type I error
Type II error

As with any two-way decision process, we can make an error by falsely rejecting the null hypothesis or by falsely accepting the null hypothesis. We give these errors the special names **Type I error** and **Type II error.**

DEFINITION 5.1

A **Type I error** is committed if we reject the null hypothesis when it is true. The probability of a Type I error is denoted by the symbol α.

DEFINITION 5.2

A **Type II error** is committed if we accept the null hypothesis when it is false and the research hypothesis is true. The probability of a Type II error is denoted by the symbol β (Greek letter beta).

The two-way decision process is shown in Table 5.3 with corresponding probabilities associated with each situation.

Although it is desirable to determine the acceptance and rejection regions to simultaneously minimize both α and β, this is not possible. The probabilities associated with Type I and Type II errors are inversely related. For a fixed sample size n, as we change the rejection region to increase α, then β decreases, and vice versa.

To alleviate what appears to be an impossible bind, the experimenter specifies a tolerable probability for a Type I error of the statistical test. Thus, the experimenter may choose α to be .01, .05, .10, and so on. Specification of a value for α then locates

TABLE 5.3
Two-way decision process

	Null Hypothesis	
Decision	**True**	**False**
Reject H_0	Type I error α	Correct $1 - \beta$
Accept H_0	Correct $1 - \alpha$	Type II error β

the rejection region. Determination of the associated probability of a Type II error is more complicated and will be delayed until later in the chapter.

Let us now see how the choice of α locates the rejection region. Returning to our soybean example, we will reject the null hypothesis for large values of the sample mean $\bar{y}$. Suppose we have decided to take a sample of $n = 36$ 1-acre plots and from these data we compute $\bar{y} = 573$ and $s = 124$. Can we conclude that the mean yield for all farms is above 520?

specifying α

Before answering this question, we must **specify α.** If we are willing to take the risk that 1 time in 40 we would incorrectly reject the null hypothesis, then $\alpha = 1/40 = .025$. An appropriate rejection region can be specified for this value of α by referring to the sampling distribution of $\bar{y}$. Assuming that $\mu = 520$ and n is large enough so that σ can be replaced by s, then $\bar{y}$ is normally distributed, with $\mu = 520$ and $\sigma/\sqrt{n} \approx 124/\sqrt{36} = 20.67$. Because the shaded area of Figure 5.6(a) corresponds to α, locating a rejection region with an area of .025 in the right tail of the distribution of $\bar{y}$ is equivalent to determining the value of z that has an area .025 to its right. Referring to Table 1 in the Appendix, this value of z is 1.96. Thus, the rejection region for our example is located 1.96 standard errors ($1.96\sigma/\sqrt{n}$) above the mean $\mu = 520$. If the observed value of $\bar{y}$ is greater than 1.96 standard errors above $\mu = 520$, we reject the null hypothesis, as shown in Figure 5.6(a).

The reason that we need to consider only $\mu = 520$ in computing α is that for all other values of μ in H_0—that is, $\mu < 520$—the probability of Type I error would be smaller than the probability of Type I error when $\mu = 520$. This can be seen by examining Figure 5.6(b)

The area of the rejection region under the curve centered at 500 is less than the area of that associated with the curve centered at 520. Thus, α for $\mu = 500$ is less than α for $\mu = 520$—that is, $\alpha\,(500) < \alpha\,(520) = .025$.

This conclusion can be extended to any value of μ less than 520—that is, all values of μ in H_0: $\mu \leq 520$.

FIGURE 5.6(a)
Rejection region for the soybean example when $\alpha = .025$

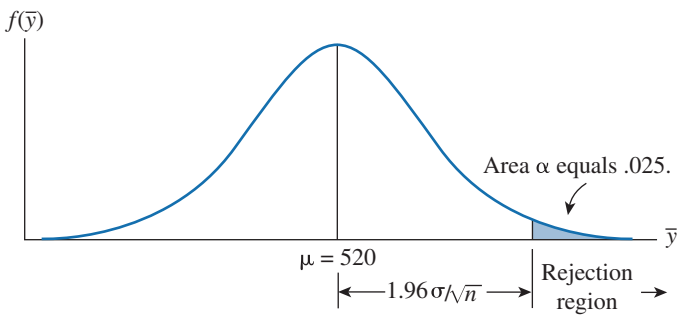
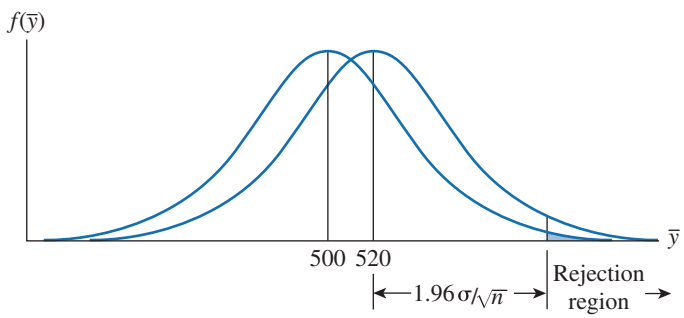


FIGURE 5.6(b)
Size of rejection region when $\mu = 500$



EXAMPLE 5.5

The Texas A&M extension service wanted to investigate if the mean yield per acre of soybeans (in bushels) was greater than 520 bushels. In a random sample of 36 1-acre soybean plots, the sample mean and standard deviation were computed to be $\bar{y} = 573$ and $s = 124$, respectively.

Set up all the parts of a statistical test for the soybean example, and use the sample data to reach a decision on whether to accept or reject the null hypothesis. Set $\alpha = .025$. Assume that σ can be estimated by s.

Solution The first four parts of the test are as follows.

H_0: $\mu \leq 520$
H_a: $\mu > 520$
T.S.: $\bar{y}$
R.R.: For $\alpha = .025$, reject the null hypothesis if $\bar{y}$ lies more than 1.96 standard errors above $\mu = 520$.

The computed value of $\bar{y}$ is 573. To determine the number of standard errors that $\bar{y}$ lies above $\mu = 520$, we compute a z-score for $\bar{y}$ using the formula

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

Substituting into the formula with s replacing σ, we have

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{573 - 520}{124/\sqrt{36}} = 2.56$$

Before drawing conclusions from these calculations, it is necessary to check the assumptions underlying the probability statements. Thus, it is necessary to make sure that the 36 1-acre soybean plots are representative of the population for which inferences are to be drawn and to examine the location of the plots to make sure that there are no confounding factors that could result in a strong correlation among the yields of the 36 plots. Finally, a normal quantile plot should be used to assess whether the 36 yields appear to be a random sample from a population having a normal distribution. Because the observed value of $\bar{y}$ lies more than 1.96—in fact it is 2.56—standard errors above 520, we reject the null hypothesis in favor of the research hypothesis and conclude that there is strong evidence in the data that average soybean yield per acre is greater than 520 bushels. ■

one-tailed test

The statistical test conducted in Example 5.5 is called a **one-tailed test** because the rejection region is located in only one tail of the distribution of $\bar{y}$. If our research hypothesis was H_a: $\mu < 520$, small values of $\bar{y}$ would indicate rejection of the null hypothesis. This test would also be one-tailed, but the rejection region would be located in the lower tail of the distribution of $\bar{y}$. Figure 5.7 displays the rejection region for the alternative hypothesis H_a: $\mu < 520$ when $\alpha = .025$.

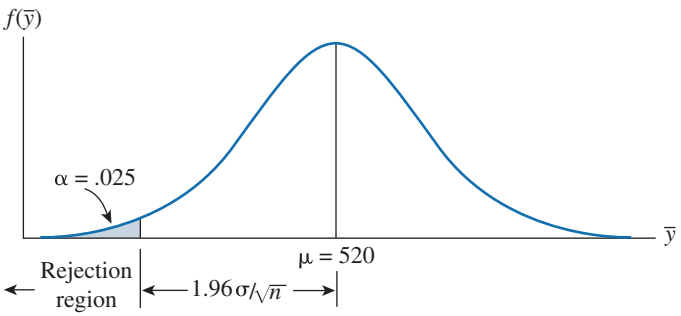
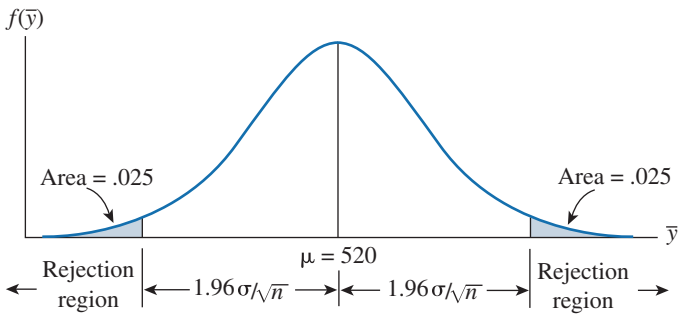


FIGURE 5.7
Rejection region for H_a: $\mu < 520$ when $\alpha = .025$ for the soybean example

FIGURE 5.8
Two-tailed rejection region for H_a: $\mu \neq 520$ when $\alpha = .05$ for the soybean example



two-tailed test

We can formulate a **two-tailed test** for the research hypothesis H_a: $\mu \neq 520$, where we are interested in detecting whether the mean yield per acre of soybeans is different from 520. Clearly, both large and small values of $\bar{y}$ would contradict the null hypothesis, and we would locate the rejection region in both tails of the distribution of $\bar{y}$. A two-tailed rejection region for H_a: $\mu \neq 520$ and $\alpha = .05$ is shown in Figure 5.8.

EXAMPLE 5.6

Elevated serum cholesterol levels are often associated with cardiovascular disease. Cholesterol levels are often thought to be associated with type of diet, amount of exercise, and genetically related factors. A recent study examined cholesterol levels among recent immigrants from China. Researchers did not have any prior information about these people and wanted to evaluate whether their mean cholesterol level differed from the mean cholesterol level of middle-aged women in the United States. The distribution of cholesterol levels in U.S. women aged 30–50 is known to be approximately normally distributed with a mean of 190 mg/dL. A random sample of $n = 100$ female Chinese immigrants aged 30–50 who had immigrated to the United States in the past year was selected from USCIS records. They were administered blood tests that yielded cholesterol levels having a mean of 178.2 mg/dL and a standard deviation of 45.3 mg/dL. Is there significant evidence in the data to demonstrate that the mean cholesterol level of the new immigrants differs from 190 mg/dL?

Solution The researchers were interested in determining if the mean cholesterol level was different from 190; thus, the research hypothesis for the statistical test is H_a: $\mu \neq 190$. The null hypothesis is the negation of the research hypothesis: H_0: $\mu = 190$. With a sample size of $n = 100$, the Central Limit Theorem should hold, and, hence, the sampling distribution of $\bar{y}$ is approximately normal. Using $\alpha = .05$, $z_{\alpha/2} = z_{.025} = 1.96$. The two-tailed rejection region for this test is given by

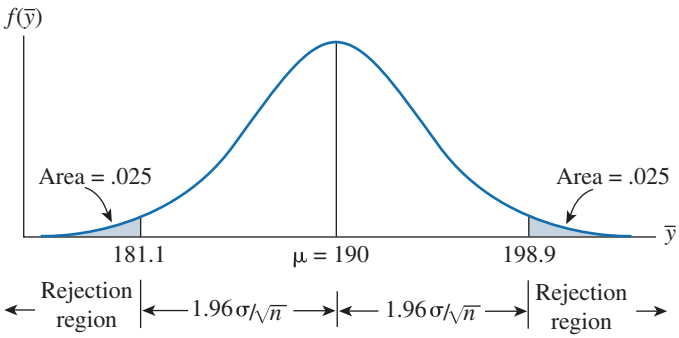
$$\mu_0 \pm 1.96s/\sqrt{n} = 190 \pm 1.96(45.3)/\sqrt{100} = 190 \pm 8.88$$

$$\text{lower rejection} = 181.1 \quad \text{upper rejection} = 198.9$$

The two regions are shown in Figure 5.9.

We can observe from Figure 5.9 that $\bar{y} = 178.2$ falls into the lower rejection region. Therefore, we conclude there is significant evidence in the data that the mean cholesterol level of middle-aged Chinese immigrants differs from 190 mg/dL.

FIGURE 5.9
Rejection region for H_a: $\mu \neq 190$ when $\alpha = .05$



Alternatively, we can determine how many standard errors $\bar{y}$ lies away from $\mu = 190$ and compare this value to $z_{\alpha/2} = z_{.025} = 1.96$. From the data, we compute

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{178.2 - 190}{45.3/\sqrt{100}} = -2.60$$

The observed value for $\bar{y}$ lies more than 1.96 standard errors below the specified mean value of 190, so we reject the null hypothesis in favor of the alternative H_a: $\mu \neq 190$. We have thus reached the same conclusion as we reached using the rejection region. The two methods will always result in the same conclusion. ■

The mechanics of the statistical test for a population mean can be greatly simplified if we use z rather than $\bar{y}$ as a test statistic. Using

H_0: $\mu \leq \mu_0$ (where μ_0 is some specified value)
H_a: $\mu > \mu_0$

and the test statistic

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

then for $\alpha = .025$ we reject the null hypothesis if $z \geq 1.96$—that is, if $\bar{y}$ lies more than 1.96 standard errors above the mean. Similarly, for $\alpha = .05$ and H_a: $\mu \neq \mu_0$, we reject the null hypothesis if the computed value of $z \geq 1.96$ or the computed value of $z \leq -1.96$. This is equivalent to rejecting the null hypothesis if the computed value of $|z| \geq 1.96$.

test for a population mean

The statistical **test for a population mean** μ is summarized next. Three different sets of hypotheses are given with their corresponding rejection regions. In a given situation, you will choose only one of the three alternatives with its associated rejection region. The tests given are appropriate only when the population distribution is normal with known σ. The rejection region will be approximately the correct region even when the population distribution is nonnormal provided the sample size is large. We can then apply the results from the Central Limit Theorem with the sample standard deviation s replacing σ to conclude that the sampling distribution of $z = (\bar{y} - \mu_0)/(s/\sqrt{n})$ is approximately normal.

Summary of a Statistical Test for μ with a Normal Population Distribution (σ Known) or Large Sample Size n

Hypotheses:

Case 1. H_0: $\mu \leq \mu_0$ vs. H_a: $\mu > \mu_0$ (right-tailed test)
Case 2. H_0: $\mu \geq \mu_0$ vs. H_a: $\mu < \mu_0$ (left-tailed test)
Case 3. H_0: $\mu = \mu_0$ vs. H_a: $\mu \neq \mu_0$ (two-tailed test)

T.S.: $z = \dfrac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$

R.R.: For a probability α of a Type I error,

Case 1. Reject H_0 if $z \geq z_\alpha$.
Case 2. Reject H_0 if $z \leq -z_\alpha$.
Case 3. Reject H_0 if $|z| \geq z_{\alpha/2}$.

Note: These procedures are appropriate if the population distribution is normally distributed with σ known. If the sample size is large, then the Central Limit Theorem allows us to use these procedures when the population distribution is nonnormal. Also, if the sample size is large, then we can replace σ with the sample standard deviation s. The situation in which n is small is presented later in this chapter.

EXAMPLE 5.7

As a part of her evaluation of municipal employees, the city manager audits the parking tickets issued by city parking officers to determine the number of tickets that were contested by the car owner and found to be improperly issued. In past years, the number of improperly issued tickets per officer had a normal distribution with mean $\mu = 380$ and standard deviation $\sigma = 35.2$. Because there has recently been a change in the city's parking regulations, the city manager suspects that the mean number of improperly issued tickets has increased. An audit of 50 randomly selected officers is conducted to test whether there has been an increase in improper tickets. Use the sample data given here and $\alpha = .01$ to test the research hypothesis that the mean number of improperly issued tickets is greater than 380. The audit generates the following data: $n = 50$ and $\bar{y} = 390$.

Solution Using the sample data with $\alpha = .01$, the five parts of a statistical test are as follows.

H_0: $\mu \leq 380$
H_a: $\mu > 380$

T.S.: $z = \dfrac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{390 - 380}{35.2/\sqrt{50}} = \dfrac{10}{35.2/7.07} = 2.01$

R.R.: For $\alpha = .01$ and a right-tailed test, we reject H_0 if $z \geq z_{.01}$, where $z_{.01} = 2.33$.

Check assumptions and draw conclusions: Because the observed value of z, 2.01, does not exceed 2.33, we might be tempted to accept the null hypothesis that $\mu \leq 380$. The only problem with this conclusion is that we do not know β, the probability of incorrectly accepting the null hypothesis. To hedge somewhat in situations in which z does not fall in the rejection region and β has not been calculated, we recommend stating that there is insufficient evidence to reject the null hypothesis. ■

computing β

We can illustrate the **computation of β,** the probability of a Type II error, using the data in Example 5.7. If the null hypothesis is H_0: $\mu \leq 380$, the probability of incorrectly accepting H_0 will depend on how close the actual mean is to 380. For example, if the actual mean number of improperly issued tickets is 400, we would expect β to be much smaller than if the actual mean is 387. The closer the actual mean is to μ_0, the more likely we are to obtain data having a value $\bar{y}$ in the acceptance region. The whole process of determining β for a test is a "what-if" type of process. In practice, we compute the value of β for a number of values of μ in the alternative hypothesis H_a and plot β versus μ in a graph called the **OC curve.** Alternatively, tests of hypotheses are evaluated by computing the probability that the test rejects false null hypotheses, called the **power** of the test. We note that power $= 1 - \beta$. The plot of power versus the value of μ is called the **power curve.** We attempt to design tests that have large values of power and hence small values for β.

OC curve

power

power curve

Let us suppose that the actual mean number of improper tickets is 395 per officer. What is β? With the null and research hypotheses as before,

$$H_0: \ \mu \leq 380$$
$$H_a: \ \mu > 380$$

and with $\alpha = .01$, we use Figure 5.10(a) to display β. The shaded portion of Figure 5.10(a) represents β, as this is the probability of $\bar{y}$ falling in the acceptance region when the null hypothesis is false and the actual value of μ is 395. The power of the test for detecting that the actual value of μ is 395 is $1 - \beta$, the area in the rejection region.

Let us consider two other possible values for μ—namely, 387 and 400. The corresponding values of β are shown as the shaded portions of Figures 5.10(b) and (c), respectively; power is the unshaded portion in the rejection region of Figures 5.10(b) and (c). The three situations illustrated in Figure 5.10 confirm what we alluded to earlier; that is, the probability of a Type II error β decreases (and hence power increases) the farther μ lies away from the hypothesized mean under H_0.

The following notation will facilitate the calculation of β. Let μ_0 denote the null value of μ, and let μ_a denote the actual value of the mean in H_a. Let $\beta(\mu_a)$ be the probability of a Type II error if the actual value of the mean is μ_a, and let PWR(μ_a) be the power at μ_a. Note that PWR(μ_a) equals $1 - \beta(\mu_a)$. Although we never really know the actual mean, we select feasible values of μ and determine β for each of these values. This will allow us to determine the probability of a Type II error occurring if one of these feasible values happens to be the actual value of the mean. The decision whether or not to accept H_0 depends on the magnitude of β for one or more reasonable values for μ_a. Alternatively, researchers calculate the power curve for a test of hypotheses. Recall that the power of the test at μ_a, PWR(μ_a), is the probability the test will detect that H_0 is false when the actual value of μ is μ_a. Hence, we want tests of hypotheses in which PWR(μ_a) is large when μ_a is in H_a and is far from μ_0.

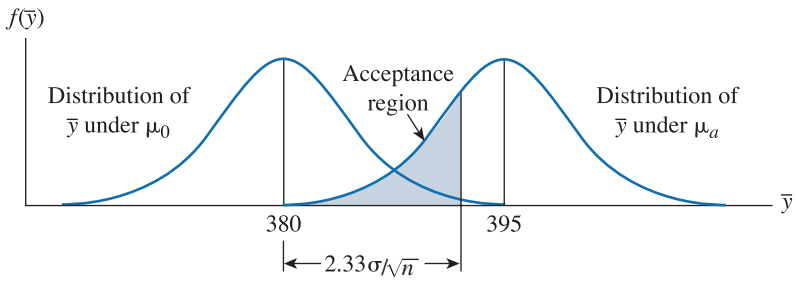
For a one-tailed test, H_0: $\mu \leq \mu_0$ or H_0: $\mu \geq \mu_0$, the value of β at μ_a is the probability that z is less than

$$z_\alpha - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}}$$

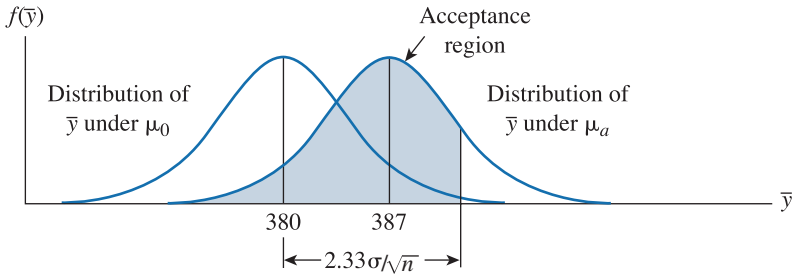
This probability is written as

$$\beta(\mu_a) = P\left[z < z_\alpha - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}}\right] = p\text{norm}\left(z_\alpha - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}}\right)$$

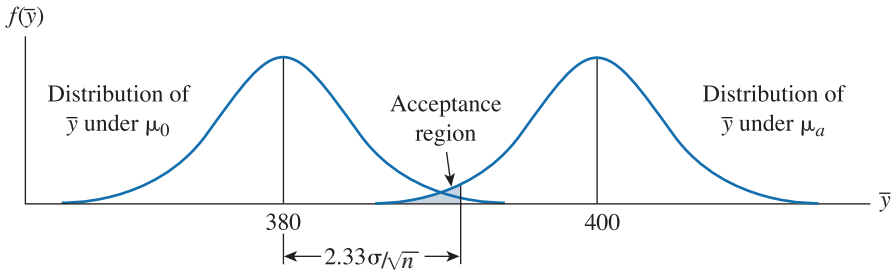
FIGURE 5.10
The probability β of a Type II error when $\mu = 395, 387$, and 400



(a) β when $\mu_a = 395$



(b) β when $\mu_a = 387$



(c) β when $\mu_a = 400$

The value of $\beta(\mu_a)$ is found by looking up the probability corresponding to the number $z_\alpha - |\mu_0 - \mu_a|/\sigma/\sqrt{n}$ in Table 1 in the Appendix.

Formulas for β are given here for one- and two-tailed tests. Examples using these formulas follow.

Calculation of β for a One- or Two-Tailed Test About μ

1. One-tailed test:

$$\beta(\mu_a) = P\left(z \le z_\alpha - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}}\right) = p\text{norm}\left(z_\alpha - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}}\right)$$

$$\text{PWR}(\mu_a) = 1 - \beta(\mu_a).$$

2. Two-tailed test:

$$\beta(\mu_a) \approx P\left(z \le z_{\alpha/2} - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}}\right) = p\text{norm}\left(z_{\alpha/2} - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}}\right)$$

$$\text{PWR}(\mu_a) = 1 - \beta(\mu_a).$$

EXAMPLE 5.8

Compute β and power for the test in Example 5.7 if the actual mean number of improperly issued tickets is 395.

Solution The research hypothesis for Example 5.7 was H_a: $\mu > 380$. Using $\alpha = .01$ and the computing formula for β with $\mu_0 = 380$ and $\mu_a = 395$, we have

$$\beta(395) = P\left[z < z_{.01} - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}}\right] = P\left[z < 2.33 - \frac{|380 - 395|}{35.2/\sqrt{50}}\right]$$

$$= P[z < 2.33 - 3.01] = P[z < -.68] = p\text{norm}(-.68) = .2483$$

Referring to Table 1 in the Appendix, the area corresponding to $z = -.68$ is .2483. Hence, $\beta(395) = .2483$ and PWR(395) $= 1 - .2483 = .7517$. ■

Previously, when $\bar{y}$ did not fall in the rejection region, we concluded that there was insufficient evidence to reject H_0 because β was unknown. Now, when $\bar{y}$ falls in the acceptance region, we can compute β corresponding to one (or more) alternative values for μ that appear reasonable in light of the experimental setting. Then, provided we are willing to tolerate a probability of falsely accepting the null hypothesis equal to the computed value of β for the alternative value(s) of μ considered, our decision is to accept the null hypothesis. Thus, in Example 5.8, if the actual mean number of improperly issued tickets is 395, then there is about a .25 probability (1 in 4 chance) of accepting the hypothesis that μ is less than or equal to 380 when in fact μ equals 395. The city manager will have to analyze the consequence of making such a decision. If the risk is acceptable, then she could state that the audit has determined that the mean number of improperly issued tickets has not increased. If the risk is too great, then the city manager will have to expand the audit by sampling more than 50 officers. In the next section, we will describe how to select the proper value for n.

EXAMPLE 5.9

As the public concern about bacterial infections increases, a soap manufacturer has quickly promoted a new product to meet the demand for an antibacterial soap. This new product has a substantially higher price than the "ordinary soaps" on the market. A consumer testing agency notes that ordinary soap also kills bacteria and questions whether the new antibacterial soap is a substantial improvement over ordinary soap. A procedure for examining the ability of soap to kill bacteria is to place a solution containing the soap onto a petri dish and then add *E. coli* bacteria. After a 24-hour incubation period, a count of the number of bacteria colonies on the dish is taken. From previous studies using many different brands of ordinary soaps, the mean bacteria count is 33 for ordinary soap products. The consumer group runs the test on the antibacterial soap using 35 petri dishes. For the 35 petri dishes, the mean bacterial count is 31.2 with a standard deviation of 8.4. Do the data provide sufficient evidence that the antibacterial soap is more effective than ordinary soap in reducing bacteria counts? Use $\alpha = .05$.

Solution Let μ be the population mean bacterial count for the antibacterial soap and σ be the population standard deviation. The five parts to our statistical test are as follows.

H_0: $\mu \geq 33$
H_a: $\mu < 33$

$$\text{T.S.:}\quad z = \frac{\bar{y} - \mu_o}{\sigma/\sqrt{n}} = \frac{31.2 - 33}{8.4/\sqrt{35}} = -1.27$$

R.R.: For $\alpha = .05$, we will reject the null hypothesis if $z \leq -z_{.05} = -1.645$.

Check assumptions and draw conclusions: With $n = 35$, the sample size is probably large enough that the Central Limit Theorem would justify our assuming that the sampling distribution of $\bar{y}$ is approximately normal. The normality assumption should be checked using the techniques from Chapter 4. Because the observed value of z, -1.27, is not less than -1.645, the test statistic does not fall in the rejection region. We reserve judgment on accepting H_0 until we calculate the chance of a Type II error, β, for several values of μ falling in the alternative hypothesis, values of μ less than 33. In other words, we conclude that there is insufficient evidence to reject the null hypothesis and hence there is not sufficient evidence that the antibacterial soap is more effective than ordinary soap. However, we next need to calculate the chance that the test may have resulted in a Type II error. ■

EXAMPLE 5.10

Refer to Example 5.9. Suppose that the consumer testing agency thinks that the manufacturer of the antibacterial soap will take legal action if the antibacterial soap has a population mean bacterial count that is considerably less than 33—say, 28. Thus, the consumer group wants to know the probability of a Type II error in its test if the population mean μ is 28 or smaller; that is, it wants to determine $\beta(28)$ because $\beta(\mu) \leq \beta(28)$ for $\mu \leq 28$.

Solution Using the computational formula for β with $\mu_0 = 33$, $\mu_a = 28$, and $\alpha = .05$, we have

$$\beta(28) = P\left[z \leq z_{.05} - \frac{|\mu_0 - \mu_a|}{\sigma/\sqrt{n}}\right] = P\left[z \leq 1.645 - \frac{|33 - 28|}{8.4/\sqrt{35}}\right]$$
$$= P[z \leq -1.88] = p\text{norm}(-1.88) = .0301$$

The area corresponding to $z = -1.88$ in Table 1 of the Appendix is .0301. Hence,

$$\beta(28) = .0301 \quad \text{and} \quad \text{PWR}(28) = 1 - .0301 = .9699$$

Because β is relatively small, we accept the null hypothesis and conclude that the antibacterial soap is not more effective than ordinary soap in reducing bacterial counts.

The manufacturer of the antibacterial soap wants to determine the chance that the consumer group may have made an error in reaching its conclusions. The manufacturer wants to compute the probability of a Type II error for a selection of potential values of μ in H_a. This would provide it with an indication of how likely it is that a Type II error may have occurred when in fact the new soap is considerably more effective in reducing bacterial counts in comparison to the mean count for ordinary soap, $\mu = 33$. Repeating the calculations for obtaining $\beta(28)$, we obtain the values in Table 5.4.

TABLE 5.4
Probability of Type II error and power for values of μ in H_a

μ	33	32	31	30	29	28	27	26	25
$\beta(\mu)$	.9500	.8266	.5935	.3200	.1206	.0301	.0049	.0005	.0000
PWR(μ)	.0500	.1734	.4065	.6800	.8794	.9699	.9951	.9995	.9999

FIGURE 5.11
Probability of Type II error

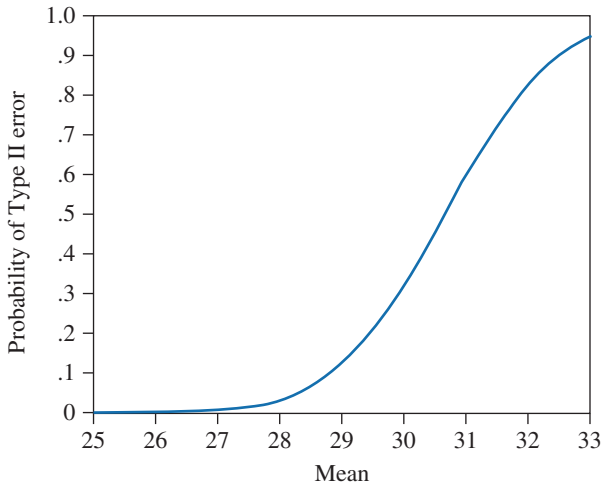
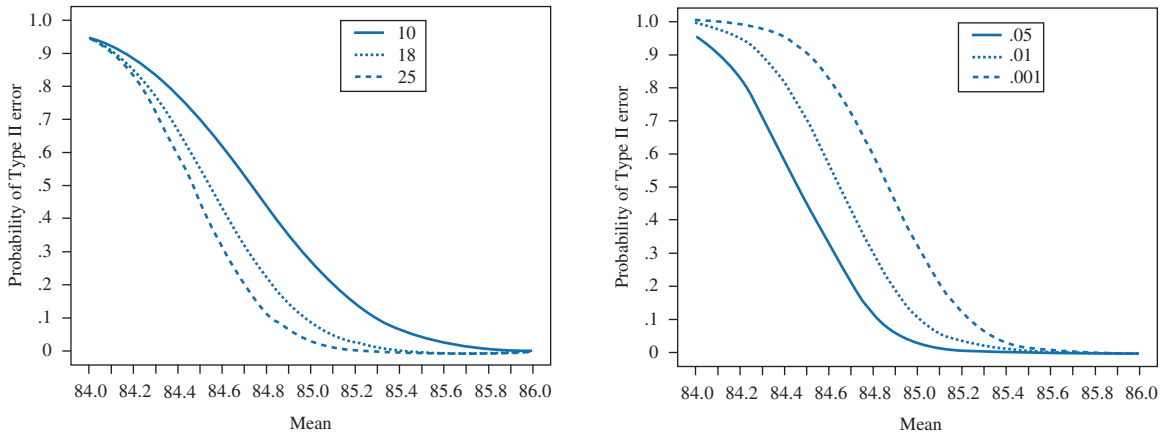


Figure 5.11 is a plot of the $\beta(\mu)$ values in Table 5.4 with a smooth curve through the points. Note that as the value of μ decreases, the probability of Type II error decreases to 0 and the corresponding power value increases to 1.0. The company could examine this curve to determine whether the chances of Type II error are reasonable for values of μ in H_a that are important to the company. From Table 5.4 or Figure 5.11, we observe that $\beta(28) = .0301$, a relatively small number. Based on the results from Example 5.9, we find that the test statistic does not fall in the rejection region. The manufacturer has decided that if the true population mean bacterial count for its antibacterial soap is 29 or less, this product is considered a substantial improvement over ordinary soap. Based on the values of the probability of Type II error displayed in Table 5.4, the chance is relatively small that the test run by the consumer agency has resulted in a Type II error for values of the mean bacterial count of 29 or smaller. Thus, the consumer testing agency was relatively certain in reporting that the new antibacterial soap did not decrease the mean bacterial count in comparison to ordinary soap. ■

In Section 5.2, we discussed how we measure the effectiveness of interval estimates. The effectiveness of a statistical test can be measured by the magnitudes of the Type I and Type II errors, α and $\beta(\mu)$. When α is preset at a tolerable level by the experimenter, $\beta(\mu_a)$ is a function of the sample size for a fixed value of μ_a. The larger the sample size n, the more information we have concerning μ, and the less likely we are to make a Type II error—hence the smaller the value of $\beta(\mu_a)$. To illustrate this idea, suppose we are testing the hypotheses H_0: $\mu \leq 84$ versus H_a: $\mu > 84$, where μ is the mean of a population having a normal distribution with $\sigma = 1.4$. If we take $\alpha = .05$, then the probability of Type II errors is plotted in Figure 5.12(a) for three possible sample sizes, $n = 10$, 18, and 25. Note that $\beta(84.6)$ becomes smaller as we increase n from 10 to 25. Another relationship of interest is that between α and $\beta(\mu)$. For a fixed sample size n, if we change the rejection region to increase the value of α, the value of $\beta(\mu_a)$ will decrease. This relationship can be observed in Figure 5.12(b). Fix the sample size at 25 and plot $\beta(\mu)$ for three different values of $\alpha = .05$, .01, and .001. We observe that $\beta(84.6)$ becomes smaller as α increases from .001 to .05. A similar set of graphs can be obtained for the power of the test by simply plotting $\text{PWR}(\mu) = 1 - \beta(\mu)$ versus μ. The relationships described would be reversed; that is, for fixed α, increasing the value of the sample size would increase the value of $\text{PWR}(\mu)$, and for fixed sample size, increasing the value of α would

FIGURE 5.12 Impact of α and n on $\beta(\mu)$



(a) $\beta(\mu)$ curve for $\alpha = .05, n = 10, 18, 25$

(b) $\beta(\mu)$ curve for $n = 25, \alpha = .05, .01, .001$

increase the value of PWR(μ). We will consider now the problem of designing an experiment for testing hypotheses about μ when α is specified and $\beta(\mu_a)$ is preset for a fixed value μ_a. This problem reduces to determining the sample size needed to achieve the fixed values of α and $\beta(\mu_a)$. Note that in those cases in which the determined value of n is too large for the initially specified values of α and β, we can increase our specified value of α and achieve the desired value of $\beta(\mu_a)$ with a smaller sample size.

5.5 Choosing the Sample Size for Testing μ

The quantity of information available for a statistical test about μ is measured by the magnitudes of the Type I and II error probabilities, α and $\beta(\mu)$, for various values of μ in the alternative hypothesis H_a. Suppose that we are interested in testing H_0: $\mu \leq \mu_0$ against the alternative H_a: $\mu > \mu_0$. First, we must specify the value of α. Next, we must determine a value of μ in the alternative, μ_1, such that if the actual value of the mean is larger than μ_1, then the consequences of making a Type II error will be substantial. Finally, we must select a value for $\beta(\mu_1)$, β. Note that for any value of μ larger than μ_1, the probability of a Type II error will be smaller than $\beta(\mu_1)$; that is,

$$\beta(\mu) < \beta(\mu_1), \text{ for all } \mu > \mu_1$$

Let $\Delta = \mu_1 - \mu_0$. The sample size necessary to meet these requirements is

$$n = \sigma^2 \frac{(z_\alpha + z_\beta)^2}{\Delta^2}$$

Note: If σ^2 is unknown, substitute an estimated value from previous studies or a pilot study to obtain an approximate sample size.

The same formula applies when testing H_0: $\mu \geq \mu_0$ against the alternative H_a: $\mu < \mu_0$, with the exception that we want the probability of a Type II error to be of magnitude β or less when the actual value of μ is *less* than μ_1, a value of the mean in H_a; that is,

$$\beta(\mu) < \beta, \text{ for all } \mu < \mu_1$$

with $\Delta = \mu_0 - \mu_1$.

EXAMPLE 5.11

A cereal manufacturer produces cereal in boxes having a labeled weight of 16 ounces. The boxes are filled by machines that are set to have a mean fill per box of 16.37 ounces. Because the actual weight of a box filled by these machines has a normal distribution with a standard deviation of approximately .225 ounces, the percentage of boxes with a fill weighing less than 16 ounces is 5% using this setting. The manufacturer is concerned that one of its machines is underfilling the boxes and wants to sample boxes from the machine's output to determine whether the mean weight μ is less than 16.37—that is, to test

$$H_0: \quad \mu \geq 16.37$$
$$H_a: \quad \mu < 16.37$$

with $\alpha = .05$. If the true mean weight is 16.27 or less, the manufacturer needs the probability of failing to detect this underfilling of the boxes with a probability of at most .01, or it risks incurring a civil penalty from state regulators. Thus, we need to determine the sample size n such that our test of H_0 versus H_a has $\alpha = .05$ and $\beta(\mu)$ less than .01 whenever μ is less than 16.27 ounces.

Solution We have $\alpha = .05$, $\beta = .01$, $\Delta = 16.37 - 16.27 = .1$, and $\sigma = .225$. Using our formula with $z_{.05} = 1.645$ and $z_{.01} = 2.33$, we have

$$n = \frac{(.225)^2(1.645 + 2.33)^2}{(.1)^2} = 79.99 \approx 80$$

Thus, the manufacturer must obtain a random sample of $n = 80$ boxes to conduct this test under the specified conditions.

Suppose that after obtaining the sample, we compute $\bar{y} = 16.35$ ounces. The computed value of the test statistic is

$$z = \frac{\bar{y} - 16.37}{\sigma/\sqrt{n}} = \frac{16.35 - 16.37}{.225/\sqrt{80}} = -.795$$

Because the rejection region is $z < -1.645$, the computed value of z does not fall in the rejection region. What is our conclusion? Knowing that $\beta(\mu) \leq .01$ when $\mu \leq 16.27$, the manufacturer is somewhat secure in concluding that the mean fill from the examined machine is at least 16.37 ounces. ■

With a slight modification of the sample size formula for the one-tailed tests, we can test

$$H_0: \quad \mu = \mu_0$$
$$H_a: \quad \mu \neq \mu_0$$

for a specified α, β, and Δ, where

$$\beta(\mu) \leq \beta, \text{ whenever } |\mu - \mu_0| \geq \Delta$$

Thus, the probability of Type II error is at most β whenever the actual mean differs from μ_0 by at least Δ. A formula for an approximate sample size n when testing a two-sided hypothesis for μ is presented here:

Approximate Sample Size for a Two-Sided Test of H_0: $\mu = \mu_0$

$$n \cong \frac{\sigma^2}{\Delta^2}(z_{\alpha/2} + z_\beta)^2$$

Note: If σ^2 is unknown, substitute an estimated value to get an approximate sample size.

5.6 The Level of Significance of a Statistical Test

In Section 5.4, we introduced hypothesis testing along rather traditional lines: We defined the parts of a statistical test along with the two types of errors, α and $\beta(\mu_a)$, and their associated probabilities. The problem with this approach is that if other researchers want to apply the results of your study using a different value for α, then they must compute a new rejection region before reaching a decision concerning H_0 and H_a. An alternative approach to hypothesis testing contains the following steps: Specify the null and alternative hypotheses, specify a value for α, collect the sample data, and determine the weight of evidence for rejecting the null hypothesis. This weight, given in terms of a probability, is called the **level of significance** (or ***p*-value**) of the statistical test. More formally, the level of significance is defined as follows: *the probability of obtaining a value of the test statistic that is as likely or more likely to reject H_0 as the actual observed value of the test statistic, assuming that the null hypothesis is true.* Thus, if the level of significance is a small value, then the sample data fail to support H_0, and our decision is to reject H_0. On the other hand, if the level of significance is a large value, then we fail to reject H_0. We must next decide what is a large or small value for the level of significance. The following decision rule yields results that will always agree with the testing procedures we introduced in Section 5.5.

level of significance
***p*-value**

Decision Rule for Hypothesis Testing Using the *p*-Value

1. If the p-value $\leq \alpha$, then reject H_0.
2. If the p-value $> \alpha$, then fail to reject H_0.

We illustrate the calculation of a level of significance with several examples.

EXAMPLE 5.12

Refer to Example 5.7.

a. Determine the level of significance (p-value) for the statistical test, and reach a decision concerning the research hypothesis using $\alpha = .01$.
b. If the preset value of α is .05 instead of .01, does your decision concerning H_a change?

Solution

a. The null and alternative hypotheses are

$$H_0: \quad \mu \leq 380$$
$$H_a: \quad \mu > 380$$

From the sample data, with s replacing σ, the computed value of the test statistic is

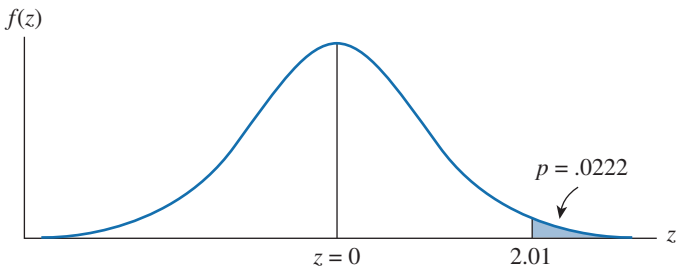
$$z = \frac{\bar{y} - 380}{\sigma/\sqrt{n}} = \frac{390 - 380}{35.2/\sqrt{50}} = 2.01$$

The level of significance for this test (i.e., the weight of evidence for rejecting H_0) is the probability of observing a value of $\bar{y}$ greater than or equal to 390 assuming that the null hypothesis is true; that is, $\mu = 380$. This value can be computed by using the z-value of the test statistic, 2.01, because

$$\begin{aligned} p\text{-value} &= P(\bar{y} \geq 390, \text{ assuming } \mu = 380) = P(z \geq 2.01) \\ &= 1 - p\text{norm}(2.01) = .0222 \end{aligned}$$

Referring to Table 1 in the Appendix, $P(z \geq 2.01) = 1 - P(z < 2.01) = 1 - .9778 = .0222$. This value is shown by the shaded area in Figure 5.13. Because the p-value is greater than α ($.0222 > .01$), we fail to reject H_0 and conclude that the data do not support the research hypothesis.

FIGURE 5.13
Level of significance for Example 5.12



b. Another person examines the same data but with a preset value for $\alpha = .05$. This person is willing to support a higher risk of a Type I error, and, hence, the decision is to reject H_0 because the p-value is less than α ($.0222 \leq .05$). It is important to emphasize that the value of α used in the decision rule is *preset* and not selected after calculating the p-value. ■

As we can see from Example 5.12, the level of significance represents the probability of observing a sample outcome more contradictory to H_0 than the observed sample result. *The smaller the value of this probability, the heavier the weight of the sample evidence against* H_0. For example, a statistical test with a level of significance of $p = .01$ shows more evidence for the rejection of H_0 than does another statistical test with $p = .20$.

EXAMPLE 5.13

Refer to Example 5.9. Using a preset value of $\alpha = .05$, is there sufficient evidence in the data to support the research hypothesis?

Solution The null and alternative hypotheses are

H_0: $\mu \geq 33$
H_a: $\mu < 33$

From the sample data, with s replacing σ, the computed value of the test statistic is

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{31.2 - 33}{8.4/\sqrt{35}} = -1.27$$

The level of significance for this test statistic is computed by determining which values of $\bar{y}$ are more extreme to H_0 than the observed $\bar{y}$. Because H_a specifies μ less than 33, the values of $\bar{y}$ that would be more extreme to H_0 are those values less than 31.2, the observed value. Thus,

$$p\text{-value} = P(\bar{y} \leq 31.2, \text{assuming } \mu = 33) = P(z \leq -1.27) = .1020$$

There is considerable evidence to support H_0. More precisely, p-value $= .1020 > .05 = \alpha$, and, hence, we fail to reject H_0. Thus, we conclude that there is insufficient evidence (p-value $= .1020$) to support the research hypothesis. Note that this is exactly the same conclusion reached using the traditional approach. ■

For two-tailed tests, H_a: $\mu \neq \mu_0$, we still determine the level of significance by computing the probability of obtaining a sample having a value of the test statistic that is more contradictory to H_0 than the observed value of the test statistic. However, for two-tailed research hypotheses, we compute this probability in terms of the magnitude of the distance from $\bar{y}$ to the null value of μ because both values of $\bar{y}$ much less than μ_0 and values of $\bar{y}$ much larger than μ_0 contradict $\mu = \mu_0$. Thus, the level of significance is written as

$$\begin{aligned} p\text{-value} &= P(|\bar{y} - \mu_0| \geq \text{observed } |\bar{y} - \mu_0|) = P(|z| \geq |\text{computed } z|) \\ &= 2P(z \geq |\text{computed } z|) \end{aligned}$$

To summarize, the level of significance (p-value) can be computed as

Case 1	Case 2	Case 3
H_0: $\mu \leq \mu_0$	H_0: $\mu \geq \mu_0$	H_0: $\mu = \mu_0$
H_a: $\mu > \mu_0$	H_a: $\mu < \mu_0$	H_a: $\mu \neq \mu_0$
p-value: $P(z \geq \text{computed } z)$	$P(z \leq \text{computed } z)$	$2P(z \geq \lvert\text{computed } z\rvert)$

EXAMPLE 5.14

Refer to Example 5.6. Using a preset value of $\alpha = .01$, is there sufficient evidence in the data to support the research hypothesis?

Solution The null and alternative hypotheses are

$$\begin{aligned} H_0&: \quad \mu = 190 \\ H_a&: \quad \mu \neq 190 \end{aligned}$$

From the sample data, with s replacing σ, the computed value of the test statistic is

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{178.2 - 190}{45.3/\sqrt{100}} = -2.60$$

The level of significance for this test statistic is computed using the formula given in Example 5.13, Case 3.

$$\begin{aligned} p\text{-value} &= 2P(z \geq |\,\text{computed } z|) = 2P(z \geq |-2.60|) = 2P(z \geq 2.60) \\ &= 2(1 - .9953) = .0094 \end{aligned}$$

Because the p-value is very small, there is very little evidence to support H_0. More precisely, p-value $= .0094 \leq .01 = \alpha$, and, hence, we reject H_0. Thus, there is sufficient evidence (p-value $= .0094$) to support the research hypothesis and conclude that the mean cholesterol level differs from 190. Note that this is exactly the same conclusion reached using the traditional approach. ■

There is much to be said in favor of this approach to hypothesis testing. Rather than reaching a decision directly, the statistician (or person performing the statistical test) presents the experimenter with the weight of evidence for rejecting the null hypothesis. The experimenter can then draw his or her own conclusion. Some experimenters reject a null hypothesis if $p \leq .10$, whereas others require $p \leq .05$ or $p \leq .01$ for rejecting the null hypothesis. The experimenter is left to make the decision based on what he or she believes is enough evidence to indicate rejection of the null hypothesis.

Many professional journals have followed this approach by reporting the results of a statistical test in terms of its level of significance. Thus, we might read that a particular test was significant at the $p = .05$ level or perhaps the $p < .01$ level. By reporting results this way, the reader is left to draw his or her own conclusion.

One word of warning is needed here. The p-value of .05 has become a magic level, and many seem to feel that a particular null hypothesis should not be rejected unless the test achieves the .05 level or lower. This has resulted in part from the decision-based approach with α preset at .05. Try not to fall into this trap when reading journal articles or reporting the results of your statistical tests. After all, statistical significance at a particular level does not dictate importance or practical significance. Rather, it means that a null hypothesis can be rejected with a specified low risk of error. For example, suppose that a company is interested in determining whether the average number of miles driven per car per month for the sales force has risen above 2,600. Sample data from 400 cars show that $\bar{y} = 2{,}640$ and $s = 35$. For these data, the z statistic for H_0: $\mu = 2{,}600$ is $z = 22.86$ based on $\sigma = 35$; the level of significance is $p < .0000000001$. Thus, even though there has been only a 1.5% increase in the average monthly miles driven for each car, the result is (highly) statistically significant. Is this increase of any practical significance? Probably not. What we have done is proved *conclusively* that the mean μ has increased slightly.

The company should not examine just the size of the p-value. It is very important to also determine the size of the difference between the null value of the population mean μ_0 and the estimated value of the population mean $\bar{y}$. This difference is called the estimated *effect size.* In this example, the estimated effect size would be $\bar{y} - \mu_0 = 2{,}640 - 2{,}600 = 40$ miles driven per month. This is the quantity that the company should consider when attempting to determine if the change in the population mean has practical significance.

Throughout the text, we will conduct statistical tests from both the decision-based approach and the level-of-significance approach to familiarize you with both avenues of thought. For either approach, remember to consider the practical significance of your findings after drawing conclusions based on the statistical test.

5.7 Inferences About μ for a Normal Population, σ Unknown

The estimation and test procedures about μ presented earlier in this chapter were based on the assumption that the population variance was known or that we had enough observations to allow s to be a reasonable estimate of σ. In this section, we present a test that can be applied when σ is unknown, no matter what the sample size, provided the population distribution is approximately normal. In Section 5.8, we will provide inference techniques for the situation where the population distribution is nonnormal. Consider the following example. Researchers would like to

determine the average concentration of a drug in the bloodstream 1 hour after it is given to patients suffering from a rare disease. For this situation, it might be impossible to obtain a random sample of 30 or more observations at a given time. What test procedure could be used in order to make inferences about μ?

W. S. Gosset faced a similar problem around the turn of the nineteenth century. As a chemist for Guinness Breweries, he was asked to make judgments on the mean quality of various brews, but he was not supplied with large sample sizes to reach his conclusions.

Gosset thought that when he used the test statistic

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

with σ replaced by s for small sample sizes, he was falsely rejecting the null hypothesis H_0: $\mu = \mu_0$ at a slightly higher rate than that specified by α. This problem intrigued him, and he set out to derive the distribution and percentage points of the test statistic

$$\frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

for $n < 30$.

For example, suppose an experimenter sets α at a nominal level—say, .05. Then he or she expects falsely to reject the null hypothesis approximately 1 time in 20. However, Gosset proved that the actual probability of a Type I error for this test was somewhat higher than the nominal level designated by α. He published the results of his study under the pen name Student because at that time it was against company policy for him to publish his results in his own name. The quantity

$$\frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

is called the t statistic, and its distribution is called the *Student's t distribution,* or simply **Student's *t***. (See Figure 5.14.)

Student's *t*

Although the quantity

$$\frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

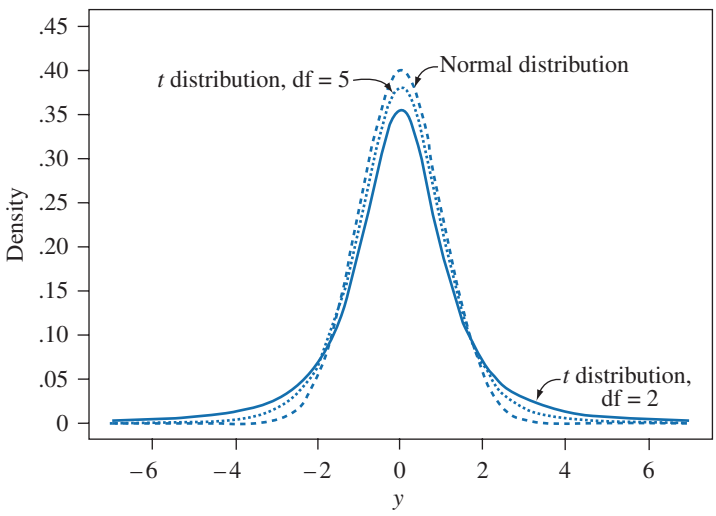


FIGURE 5.14
Two t distributions and a standard normal distribution

possesses a t distribution only when the sample is selected from a normal population, the t distribution provides a reasonable approximation to the distribution of

$$\frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

when the sample is selected from a population with a mound-shaped distribution. We summarize the properties of t here.

Properties of Student's *t* Distribution

1. There are many different t distributions. We specify a particular one by a parameter called the degrees of freedom (df). (See Figure 5.14.)
2. The t distribution is symmetrical about 0 and hence has a mean equal to 0, the same as the z distribution.
3. The t distribution has variance df/(df − 2) and hence is more variable than the z distribution, which has a variance equal to 1. (See Figure 5.14.)
4. As the df increase, the t distribution approaches the z distribution. (Note that as the df increase, the variance df/(df − 2) approaches 1.)
5. Thus, with

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

we conclude that t has a t distribution with df $= n - 1$, and as n increases, the distribution of t approaches the distribution of z.

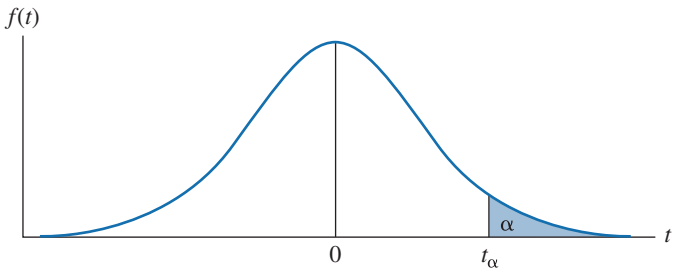
The phrase "degrees of freedom" sounds mysterious now, but the idea will eventually become second nature to you. The technical definition requires advanced mathematics, which we will avoid; on a less technical level, the basic idea is that degrees of freedom are pieces of information for estimating σ using s. The standard deviation s for a sample of n measurements is based on the deviations $y_i - \bar{y}$. Because $\Sigma(y_i - \bar{y}) = 0$ always, if $n - 1$ of the deviations are known, the last (nth) is fixed mathematically to make the sum equal 0. It is therefore noninformative. Thus, in a sample of n measurements, there are $n - 1$ pieces of information (degrees of freedom) about σ. A second method of explaining degrees of freedom is to recall that σ measures the dispersion of the population values about μ, so prior to estimating σ we must first estimate μ. Hence, the number of pieces of information (degrees of freedom) in the data that can be used to estimate σ is $n - 1$, the number of original data values minus the number of parameters estimated prior to estimating σ.

Because of the symmetry of t, only upper-tail percentage points (probabilities or areas) of the distribution of t have been tabulated; these appear in Table 2 in the Appendix. The degrees of freedom (df) are listed along the left column of the page. An entry in the table specifies a value of t—say, t_α—such that an area α lies to its right. See Figure 5.15. Various values of α appear across the top of Table 2 in the Appendix. Thus, for example, with df = 7, the value of t with an area .05 to its right is 1.895 (found in the $\alpha = .05$ column and df = 7 row). Since the t distribution approaches the z distribution as df approach ∞, the values in the last row of Table 2 are the same as z_α. Thus, we can quickly determine z_α by using values in the last row of Table 2 in the Appendix.

t_α

We can use the t distribution to make inferences about a population mean μ. The sample test concerning μ is summarized next. The only difference between the z test discussed earlier in this chapter and the test given here is that s replaces σ. The t test (rather than the z test) should be used any time σ is unknown and the distribution of y-values is mound-shaped.

FIGURE 5.15
Illustration of area tabulated in Table 2 in the Appendix for the t distribution



Summary of a Statistical Test for μ with a Normal Population Distribution (σ Unknown)

Hypotheses:

Case 1. H_0: $\mu \leq \mu_0$ vs. H_a: $\mu > \mu_0$ (right-tailed test)
Case 2. H_0: $\mu \geq \mu_0$ vs. H_a: $\mu < \mu_0$ (left-tailed test)
Case 3. H_0: $\mu = \mu_0$ vs. H_a: $\mu \neq \mu_0$ (two-tailed test)

T.S.: $t = \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}}$

R.R.: For a probability α of a Type I error and df $= n - 1$:

Case 1. Reject H_0 if $t \geq t_\alpha = qt(1 - \alpha, n - 1)$
Case 2. Reject H_0 if $t \leq -t_\alpha = -qt(1 - \alpha, n - 1)$
Case 3. Reject H_0 if $|t| \geq t_{\alpha/2} = qt(1 - \alpha/2, n - 1)$

Level of significance (p-value):

Case 1. p-value $= P(t \geq$ computed $t)$
Case 2. p-value $= P(t \leq$ computed $t)$
Case 3. p-value $= 2P(t \geq |$computed $t|)$

Recall that α denotes the area in the tail of the t distribution. For a one-tailed test with the probability of a Type I error equal to α, we locate the rejection region using the value from Table 2 in the Appendix for the specified α and df $= n - 1$. However, for a two-tailed test, we use the t-value from Table 2 corresponding to $\alpha/2$ and df $= n - 1$.

Thus, for a one-tailed test, we reject the null hypothesis if the computed value of t is greater than the t-value from Table 2 in the Appendix with the specified α and df $= n - 1$. Similarly, for a two-tailed test, we reject the null hypothesis if $|t|$ is greater than the t-value from Table 2 with $\alpha/2$ and df $= n - 1$.

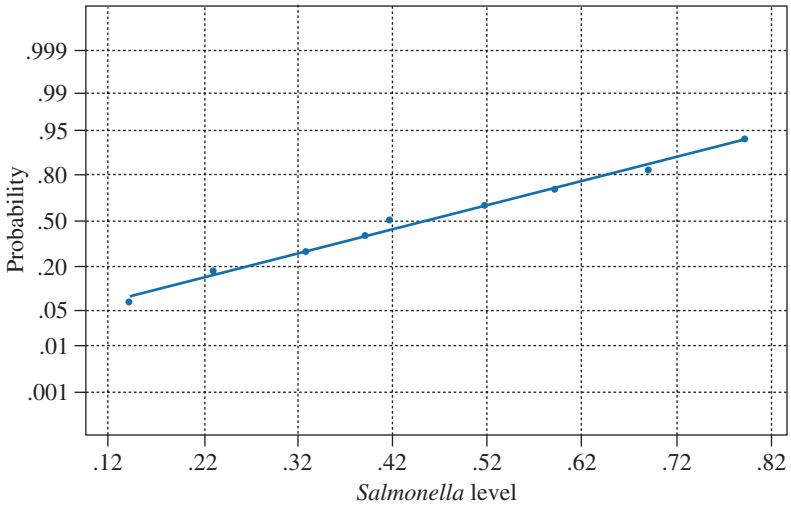
EXAMPLE 5.15

A massive multistate outbreak of foodborne illness was attributed to *Salmonella enteritidis*. Epidemiologists determined that the source of the illness was ice cream. They sampled nine production runs from the company that had produced the ice cream to determine the level of *Salmonella enteritidis* in the ice cream. These levels (MPN/g) are as follows:

.593 .142 .329 .691 .231 .793 .519 .392 .418

Use these data to determine whether the average level of *Salmonella enteritidis* in the ice cream is greater than .3 MPN/g, a level that is considered to be very dangerous. Set $\alpha = .01$.

FIGURE 5.16
Normal probability plot for *Salmonella* data



Solution The null and research hypotheses for this example are

$$H_0: \quad \mu \leq .3$$
$$H_a: \quad \mu > .3$$

Because the sample size is small, we need to examine whether the data appear to have been randomly sampled from a normal distribution. Figure 5.16 is a normal probability plot of the data values. All nine points fall nearly on the straight line. We conclude that the normality condition appears to be satisfied. Before setting up the rejection region and computing the value of the test statistic, we must first compute the sample mean and standard deviation. You can verify that

$$\bar{y} = .456 \text{ and } s = .2128$$

The rejection region with $\alpha = .01$ is

$$\text{R.R.:} \quad \text{Reject } H_0 \text{ if } t > 2.896$$

where, from Table 2 in the Appendix, the value of $t_{.01}$ with df $= 9 - 1 = 8$ is 2.896. The computed value of t is

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{.456 - .3}{.2128/\sqrt{9}} = 2.20$$

The observed value of t is not greater than 2.896, so we have insufficient evidence to indicate that the average level of *Salmonella enteritidis* in the ice cream is greater than .3 MPN/g. The level of significance of the test is given by

$$p\text{-value} = P(t > \text{computed } t) = P(t > 2.20) = 1 - pt(2.2, 8) = .029$$

Using the t-tables there are only a few areas (α) for each value of df. The best we can do is bound the p-value. From Table 2 with df $= 8, t_{.05} = 1.860$ and $t_{.025} = 2.306$. Because computed $t = 2.20$, $.025 < p\text{-value} < .05$. However, with $\alpha = .01 < .025 < p$-value, we can still conclude that p-value $> \alpha$ and hence fail to reject H_0.

In order to assess the chance of a Type II error, we need to calculate the probability of a Type II error for some crucial values of μ in H_a. These calculations are somewhat more complex than the calculations for the z test. We will use a set of graphs to determine $\beta(\mu_a)$. The value of $\beta(\mu_a)$ depends on three quantities, df $= n - 1$, α, and the distance d from μ_a to μ_0 in σ units:

$$d = \frac{|\mu_a - \mu_0|}{\sigma}$$

Thus, to determine $\beta(\mu_a)$, we must specify α and μ_a and provide an estimate of σ. Then with the calculated d and df $= n - 1$, we locate $\beta(\mu_a)$ on the graph. Table 3 in the Appendix provides graphs of $\beta(\mu_a)$ for $\alpha = .01$ and .05 for both one-sided and two-sided hypotheses for a variety of values for d and df. ■

EXAMPLE 5.16

Refer to Example 5.15. We have $n = 9$, $\alpha = .01$, and a one-sided test. Thus, df $= 8$, and if we estimate $\sigma \approx .25$, we can compute the values of d corresponding to selected values of μ_a. The values of $\beta(\mu_a)$ can then be determined using the graphs in Table 3 in the Appendix. Figure 5.17 is the necessary graph for this example. To illustrate the calculations, let $\mu_a = .45$. Then

$$d = \frac{|\mu_a - \mu_0|}{\sigma} = \frac{|.45 - .3|}{.25} = .6$$

We draw a vertical line from $d = .6$ on the horizontal axis to the curve labeled 8, our df. We then locate the value on the vertical axis at the height of the intersection, .79. Thus, $\beta(.45) = .79$. Similarly, to determine $\beta(.55)$, first compute $d = 1.0$, draw a vertical line from $d = 1.0$ to the curve labeled 8, and locate .43 on the vertical axis.

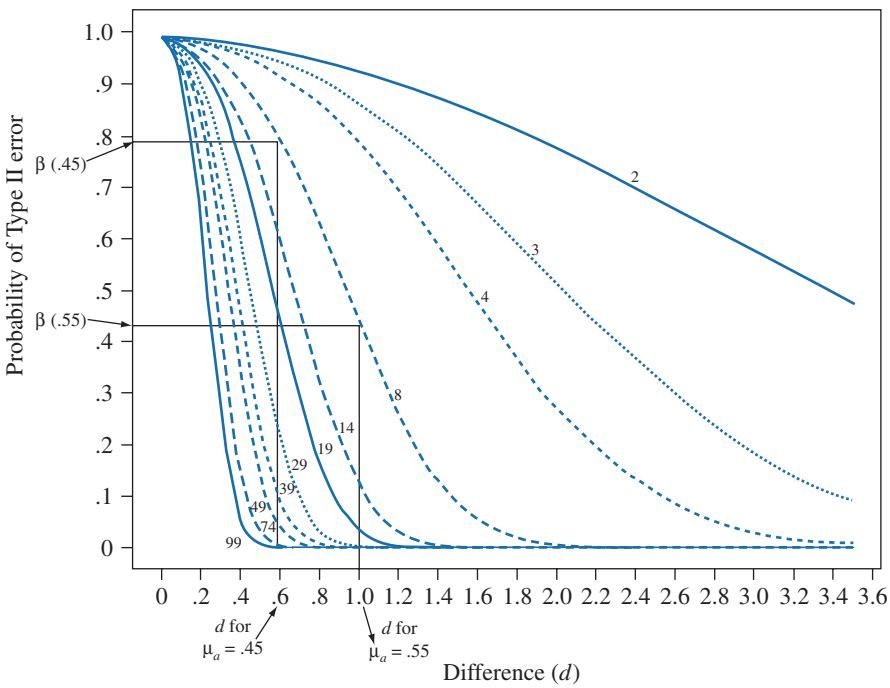


FIGURE 5.17
Probability of Type II error curves $\alpha = .01$, one-sided

Thus, $\beta(.55) = .43$. Table 5.5 contains values of $\beta(\mu_a)$ for several values of μ_a. Because the values of $\beta(\mu_a)$ are large for values of μ_a that are considerably larger than $\mu_0 = .3$—for example, $\beta(.6) = .26$—we will not state that μ is less than or equal to .3 but will only state that the data fail to support the contention that μ is larger than .3.

TABLE 5.5
Probability of Type II errors

μ_a	.35	.4	.45	.5	.55	.6	.65	.7	.75	.8
d	.2	.4	.6	.8	1.0	1.2	1.4	1.6	1.8	2.0
$\beta(\mu_a)$	.97	.91	.79	.63	.43	.26	.13	.05	.02	.00

In addition to being able to run a statistical test for μ when σ is unknown, we can construct a confidence interval using t. The confidence interval for μ with σ unknown is identical to the corresponding confidence interval for μ when σ is known, with z replaced by t and σ replaced by s.

100(1 − α)% Confidence Interval for μ, σ Unknown

$$\bar{y} \pm t_{\alpha/2}\, s/\sqrt{n}$$

Note: df $= n - 1$ and the confidence coefficient is $(1 - \alpha)$.

EXAMPLE 5.17

An airline wants to evaluate the depth perception of its pilots over the age of 50. A random sample of $n = 14$ airline pilots over the age of 50 is asked to judge the distance between two markers placed 20 feet apart at the opposite end of the laboratory. The sample data listed here are the pilots' errors (recorded in feet) in judging the distance.

2.7 2.4 1.9 2.6 2.4 1.9 2.3
2.2 2.5 2.3 1.8 2.5 2.0 2.2

Use the sample data to place a 95% confidence interval on μ, the average error in depth perception for the company's pilots over the age of 50.

Solution Before setting up a 95% confidence interval on μ, we must first assess the normality assumption by plotting the data in a normal probability plot or a boxplot. Figure 5.18 is a boxplot of the 14 data values. The median line is near the center of the box, the right and left whiskers are approximately the same length, and there are no outliers. The data appear to be a sample from a normal distribution. Thus, it is appropriate to construct the confidence interval based on the t distribution. You can verify that

$$\bar{y} = 2.26 \text{ and } s = .28$$

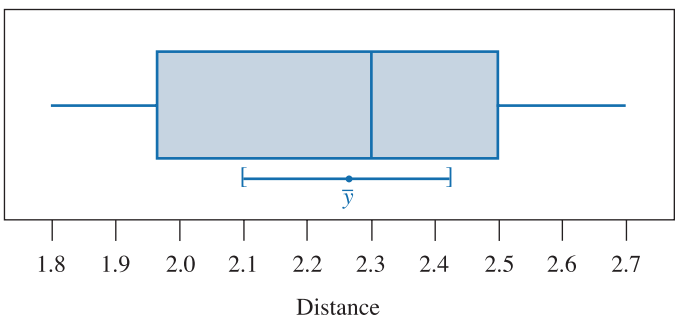


FIGURE 5.18
Boxplot of distance (with 95% t confidence interval for the mean)

Referring to Table 2 in the Appendix, the t-value corresponding to $a = .025$ and df = 13 is 2.160. Hence, the 95% confidence interval for μ is

$$\bar{y} \pm t_{\alpha/2}\, s/\sqrt{n} \quad \text{or} \quad 2.26 \pm 2.160\,(.28)/\sqrt{14}$$

which is the interval $2.26 \pm .16$, or 2.10 to 2.42. Thus, we are 95% confident that the average error in the pilots' judgment of the distance is between 2.10 and 2.42 feet. ■

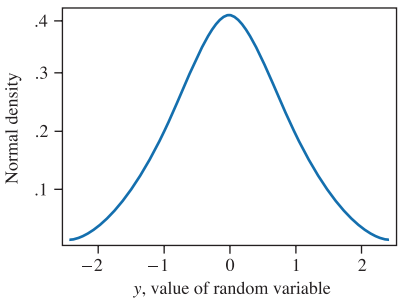
In this section, we have made the formal mathematical assumption that the population is normally distributed. *In practice, no population has exactly a normal distribution*. How does nonnormality of the population distribution affect inferences based on the t distribution?

There are two issues to consider when populations are assumed to be nonnormal. First, what kind of nonnormality is assumed? Second, what possible effects do these specific forms of nonnormality have on the t-distribution procedures? The most important deviations from normality are **skewed distributions** and **heavy-tailed distributions.** Heavy-tailed distributions are roughly symmetric but have outliers relative to a normal distribution. Figure 5.19 displays these nonnormal distributions: Figure 5.19(a) is the standard normal distribution, Figure 5.19(b) is a heavy-tailed distribution (a t distribution with df = 3), Figure 5.19(c) is a distribution mildly skewed to the right, and Figure 5.19(d) is a distribution heavily skewed to the right.

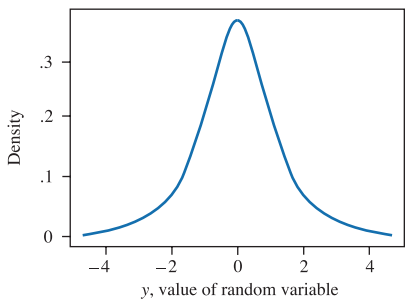
skewed distributions
heavy-tailed distributions

To evaluate the effect of nonnormality as exhibited by skewness or heavy-tailedness, we will consider whether the t-distribution procedures are still approximately correct for these forms of nonnormality and whether there are other more efficient procedures. For example, even if a test procedure for μ based on the t distribution gives nearly correct results for, say, a heavy-tailed population distribution,

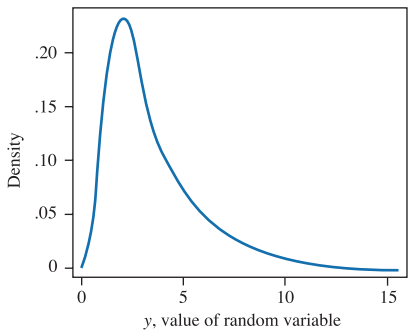
FIGURE 5.19
Standard normal distriubtion and three nonnormal distributions



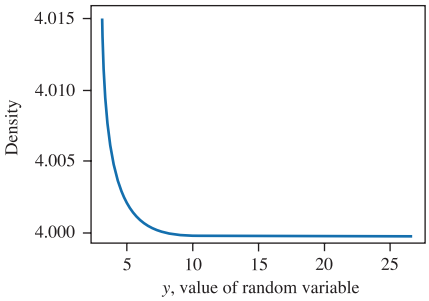
(a) Density of the standard normal distribution



(b) Density of a heavy-tailed distribution



(c) Density of a lightly skewed distribution



(d) Density of a highly skewed distribution

it might be possible to obtain a test procedure with a more accurate probability of Type I error and greater power if we test hypotheses about the population median in place of the population μ. Also, in the case of heavily tailed or highly skewed population distributions, the median rather than μ is a more appropriate representation of the population center.

The question of approximate correctness of t procedures has been studied extensively. In general, probabilities specified by the t procedures, particularly the confidence level for confidence intervals and the Type I error for statistical tests, have been found to be fairly accurate, even when the population distribution is heavy-tailed. However, when the population is very heavy-tailed, as is the case in Figure 5.19(b), the tests of hypotheses tend to have a probability of Type I errors smaller than the specified level, which leads to a test having much lower power and hence greater chances of committing Type II errors. Skewness, particularly with small sample sizes, can have an even greater effect on the probability of both Type I and Type II errors. When we are sampling from a population distribution that is normal, the sampling distribution of a t statistic is symmetric. However, when we are sampling from a population distribution that is highly skewed, the sampling distribution of a t statistic is skewed, not symmetric. Although the degree of skewness decreases as the sample size increases, there is no procedure for determining the sample size at which the sampling distribution of the t statistic becomes symmetric.

As a consequence, the level of a nominal $\alpha = .05$ test may actually have a level of .01 or less when the sample size is less than 20 and the population distribution looks like that of Figure 5.19(b), (c), or (d). Furthermore, the power of the test will be considerably less than when the population distribution is a normal distribution, thus causing an increase in the probability of Type II errors. A simulation study of the effect of skewness and heavy-tailedness on the level and power of the t test yielded the results given in Table 5.6. The values in the table are the power values for a level $\alpha = .05$ t test of H_0: $\mu \leq \mu_0$ versus H_a: $\mu > \mu_0$. The power values are calculated for shifts of size $d = |\mu_a - \mu_0|/\sigma$ for values of $d = 0, .2, .6, .8$. Three different sample sizes were used: $n = 10$, 15, and 20.When $d = 0$, the level of the test is given for each type of population distribution. We want to compare these values to .05. The values when $d > 0$ are compared to the corresponding values when sampling from a normal population. We observe that when sampling from the lightly skewed distribution and the heavy-tailed distribution, the levels are somewhat less than .05 with values nearly equal to .05 when using $n = 20$. However, when sampling from a heavily skewed distribution, even with $n = 20$ the level is only .011. The power values for the heavily tailed and heavily skewed populations are considerably less than the corresponding values when sampling from a normal distribution. Thus, the test is much less likely to correctly detect that the

TABLE 5.6
Level and power values for t test

Population Distribution	$n = 10$				$n = 15$				$n = 20$			
	Shift d				Shift d				Shift d			
	0	.2	.6	.8	0	.2	.6	.8	0	.2	.6	.8
Normal	.05	.145	.543	.754	.05	.182	.714	.903	.05	.217	.827	.964
Heavy-tailedness	.035	.104	.371	.510	.049	.115	.456	.648	.045	.163	.554	.736
Light skewness	.025	.079	.437	.672	.037	.129	.614	.864	.041	.159	.762	.935
Heavy skewness	.007	.055	.277	.463	.006	.078	.515	.733	.011	.104	.658	.873

alternative hypothesis H_a is true. This reduced power is present even when $n = 20$. When sampling from a lightly skewed population distribution, the power values are very nearly the same as the values for the normal distribution.

Because the t procedures have reduced power when sampling from skewed populations with small sample sizes, procedures have been developed that are not as affected by the skewness or extreme heavy-tailedness of the population distribution. These procedures are called **robust methods** of estimation and inference. Three robust procedures, the bootstrap, the sign test, and Wilcoxon signed rank test, will be considered in Sections 5.8 and 5.9, and Chapter 6, respectively. They are both more efficient than the t test when the population distribution is very nonnormal in shape. Also, they maintain the selected α level of the test, unlike the t test, which, when applied to very nonnormal data, has a true α value much different from the selected α value. The same comments can be made with respect to confidence intervals for the mean. When the population distribution is highly skewed, the coverage probability of a nominal $100(1 - \alpha)$ confidence interval is considerably less than $100(1 - \alpha)$.

robust methods

So what is a nonexpert to do? First, examine the data through graphs. A boxplot or normal probability plot will reveal any gross skewness or extreme outliers. If the plots do not reveal extreme skewness or many outliers, the nominal t-distribution probabilities should be reasonably correct. Thus, the level and power calculations for tests of hypotheses and the coverage probability of confidence intervals should be reasonably accurate. If the plots reveal severe skewness or heavy-tailedness, the test procedures and confidence intervals based on the t distribution will be highly suspect. In these situations, we have two alternatives. First, it may be more appropriate to consider inferences about the population median rather than the population mean. When the data are highly skewed or very heavily tailed, the median is a more appropriate measure of the center of the population than is the mean. In Section 5.9, we will develop tests of hypotheses and confidence intervals for the population median. These procedures will avoid the problems encountered by the t-based procedures discussed in this section when the population distribution is highly skewed or heavily tailed. However, in some situations, the researcher may be required to provide inferences about the mean, or the median may not be an appropriate alternative to the mean as a summary of the population. In Section 5.8, we will discuss a technique based on bootstrap methods for obtaining an approximate confidence interval for the population mean.

5.8 Inferences About μ When the Population Is Nonnormal and n Is Small: Bootstrap Methods

The statistical techniques in the previous sections for constructing a confidence interval or a test of hypotheses for μ required that the population have a normal distribution or that the sample size be reasonably large. In those situations where neither of these requirements can be met, an alternative approach using bootstrap methods can be employed. This technique was introduced by Efron in the article **"Bootstrap Methods: Another Look at the Jackknife"** ***[Annals of Statistics (1979) 7:1–26].*** The bootstrap is a technique by which an approximation to the sampling distribution of a statistic can be obtained when the population distribution is unknown. In Section 5.7, inferences about μ were based on the fact that the statistic

$$t \text{ statistic} = \frac{\bar{y} - \mu}{s/\sqrt{n}}$$

had a t distribution. We used the t-tables (Table 2 in the Appendix) to obtain appropriate percentiles and p-values for confidence intervals and tests of hypotheses. However, it was required that the population from which the sample was randomly selected have a normal distribution or that the sample size n be reasonably large. The bootstrap will provide a means for obtaining percentiles of $\frac{\bar{y} - \mu}{s/\sqrt{n}}$ when the population distribution is nonnormal and/or the sample size is relatively small.

The bootstrap technique utilizes data-based simulations for statistical inference. The central idea of the bootstrap is to resample from the original data set, thus producing a large number of replicate data sets from which the sampling distribution of a statistic can be approximated. Suppose we have a sample $y_1, y_2, \ldots, y_n$ from a population and we want to construct a confidence interval or test a set of hypotheses about the population mean μ. We realize either from prior experience with this population or from an examination of a normal quantile plot that the population has a nonnormal distribution. Thus, we are fairly certain that the sampling distribution of $t = \frac{\bar{y} - \mu}{s/\sqrt{n}}$ is not the t distribution, so it would not be appropriate to use the t-tables to obtain percentiles. Also, the sample size n is relatively small so we are not too sure about applying the Central Limit Theorem and using the z-tables to obtain percentiles to construct confidence intervals or to test hypotheses.

The bootstrap technique consists of the following steps:

1. Select a random sample $y_1, y_2, \ldots, y_n$ of size n from the population, and compute the sample mean, $\bar{y}$, and sample standard deviation, s.
2. Select a random sample of size n, with replacement from $y_1, y_2, \ldots, y_n$ yielding $y_1^*, y_2^*, \ldots, y_n^*$.
3. Compute the mean $\bar{y}^*$ and standard deviation s^* of $y_1^*, y_2^*, \ldots, y_n^*$.
4. Compute the value of the statistic
$$\hat{t} = \frac{\bar{y}^* - \bar{y}}{s^*/\sqrt{n}}$$
5. Repeat Steps 2–4 a large number of times, B, to obtain $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_B$. Use these values to obtain an approximation to the sampling distribution of $\frac{\bar{y} - \mu}{s/\sqrt{n}}$.

Suppose we have $n = 20$ and we select $B = 9{,}999$ bootstrap samples. The steps in obtaining the bootstrap approximation to the sampling distribution of $\frac{\bar{y} - \mu}{s/\sqrt{n}}$ are depicted here.

Obtain random sample $y_1, y_2, \ldots, y_{20}$ from the population, and compute $\bar{y}$ and s.

First bootstrap sample: $y_1^*, y_2^*, \ldots, y_{20}^*$ yields $\bar{y}^*$, s^*, and $\hat{t}_1 = \frac{\bar{y}^* - \bar{y}}{s^*/\sqrt{20}}$

Second bootstrap sample: $y_1^*, y_2^*, \ldots, y_{20}^*$ yields $\bar{y}^*$, s^*, and $\hat{t}_2 = \frac{\bar{y}^* - \bar{y}}{s^*/\sqrt{20}}$

.
.
.

Bth bootstrap sample: $y_1^*, y_2^*, \ldots, y_{20}^*$ yields $\bar{y}^*$, s^*, and $\hat{t}_B = \frac{\bar{y}^* - \bar{y}}{s^*/\sqrt{20}}$

We then use the B values of $\hat{t}$—$\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_B$—to obtain the approximate percentiles. For example, suppose we want to construct a 95% confidence interval for μ and $B = 9{,}999$. We need the lower and upper .025 percentiles, $\hat{t}_{.025}$ and $\hat{t}_{.975}$. Thus, we would take the $(9{,}999 + 1)(.025) = 250$th-largest value of $\hat{t} = \hat{t}_{.025}$ and the $(9{,}999 + 1)(1 - .025) = 9{,}750$th-largest value of $\hat{t} = \hat{t}_{.975}$. The approximate 95% confidence interval for μ would be

$$\left(\bar{y} - \hat{t}_{.975}\frac{s}{\sqrt{n}},\ \bar{y} - \hat{t}_{.025}\frac{s}{\sqrt{n}}\right)$$

EXAMPLE 5.18

Secondhand smoke is of great concern, especially when it involves young children. Breathing secondhand smoke can be harmful to children's health, contributing to health problems such as asthma, Sudden Infant Death Syndrome (SIDS), bronchitis and pneumonia, and ear infections. The developing lungs of young children are severely affected by exposure to secondhand smoke. Child Protective Services (CPS) in a city is concerned about the level of exposure to secondhand smoke for children placed by their agency in foster parents' care. A method of determining level of exposure is to determine the urinary concentration of cotanine, a metabolite of nicotine. Unexposed children will typically have mean cotanine levels of 75 or less. A random sample of 20 children suspected of being exposed to secondhand smoke yielded the following urinary concentrations of cotanine:

29, 30, 53, 75, 89, 34, 21, 12, 58, 84, 92, 117, 115, 119, 109, 115, 134, 253, 289, 287

CPS wants an estimate of the mean cotanine level in the children under their care. From the sample of 20 children, it computes $\bar{y} = 105.75$ and $s = 82.429$. Construct a 95% confidence interval for the mean cotanine level for children under the supervision of CPS.

Solution Because the sample size is relatively small, an assessment of whether the population has a normal distribution is crucial prior to using a confidence interval procedure based on the t distribution. Figure 5.20 displays a normal probability plot for the 20 data values. From the plot, we observe that the data do not fall near the straight line, and the p-value for the test of normality is less than .01. Thus, we would conclude that the data do not appear to follow a normal distribution. The confidence interval based on the t distribution would not be appropriate; hence, we will use a bootstrap confidence interval.

$B = 9{,}999$ samples of size 20 are selected with replacement from the original sample. Table 5.7 displays 5 of the 9,999 samples to illustrate the nature of the bootstrap samples.

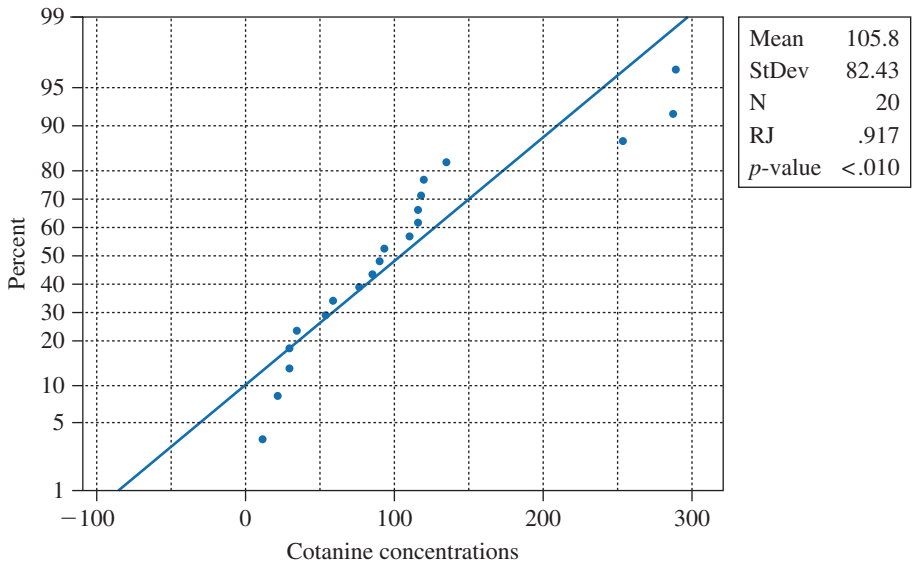


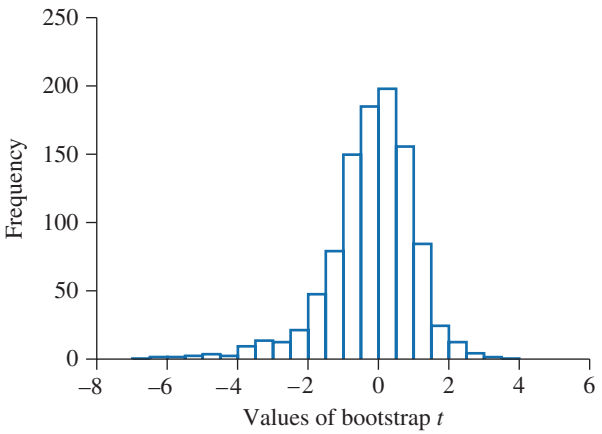
FIGURE 5.20
Normal probability plot for cotanine data

TABLE 5.7
Bootstrap samples

Original	29	30	53	75	89	34	21	12	58	84
Sample	92	117	115	119	109	115	134	253	289	287
Bootstrap	29	21	12	115	21	89	29	30	21	89
Sample 1	30	84	84	134	58	30	34	89	29	134
Bootstrap	30	92	75	109	115	117	84	89	119	289
Sample 2	115	75	21	92	109	12	289	58	92	30
Bootstrap	53	289	30	92	30	253	89	89	75	119
Sample 3	115	117	253	53	84	34	58	289	92	134
Bootstrap	75	21	115	287	119	75	75	53	34	29
Sample 4	117	115	29	115	115	253	289	134	53	75
Bootstrap	89	119	109	109	115	119	12	29	84	21
Sample 5	34	134	115	134	75	58	30	75	109	134

Upon examination of Table 5.7, it can be observed that in each of the bootstrap samples there are repetitions of some of the original data values. This arises due to the sampling with replacement. The following histogram of the 9,999 values of $\hat{t} = \frac{\bar{y}^* - \bar{y}}{s*/\sqrt{n}}$ illustrates the effect of the nonnormal nature of the population distribution on the sampling distribution on the t statistic. If the sample had been randomly selected from a normal distribution, the histogram would be symmetric, as was depicted in Figure 5.14. The histogram in Figure 5.21 is somewhat left-skewed.

FIGURE 5.21
Histogram of bootstrapped t-statistic



After sorting the 9,999 values of $\hat{t}$ from smallest to largest, we obtain the 250th-smallest and 250th-largest values: −3.167 and 1.748, respectively. We thus have the following percentiles:

$$\hat{t}_{.025} = -3.167 \quad \text{and} \quad \hat{t}_{.975} = 1.748$$

The 95% confidence interval for the mean cotanine concentration is given here using the original sample mean of $\bar{y} = 105.75$ and original sample standard deviation of $s = 82.429$:

$$\left(\bar{y} - \hat{t}_{.975}\frac{s}{\sqrt{n}},\ \bar{y} - \hat{t}_{.025}\frac{s}{\sqrt{n}}\right) = \left(105.75 - 1.748\frac{82.429}{\sqrt{20}}, 105.75 + 3.167\frac{82.429}{\sqrt{20}}\right)$$
$$= (73.53, 164.12)$$

A comparison of these two percentiles to the percentiles from the t distribution (Table 2 in the Appendix) reveals how much in error our confidence intervals would have been if we has directly applied the formulas from Section 5.7.

From Table 2 in the Appendix, with df = 19, we have $t_{.025} = -2.093$ and $t_{.975} = 2.093$. This would yield a 95% confidence interval on μ of

$$105.75 \pm 2.093 \frac{82.429}{\sqrt{20}} \Rightarrow (67.17,\ 144.33)$$

Note that the confidence interval using the t distribution is centered about the sample mean, whereas the bootstrap confidence interval has its upper limit farther from the mean than its lower limit. This is due to the fact that the random sample from the population indicated that the population distribution was not symmetric. Thus, we would expect that the sampling distribution of our statistic would not be symmetric due to the relatively small size, $n = 20$. ■

We will next apply the bootstrap approximation of the test statistic $t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$ to obtain a test of hypotheses for the situation where n is relatively small and the population distribution is nonnormal. The method for obtaining the p-value for the bootstrap approximation to the sampling distribution of the test statistic under the null value of μ, μ_0, involves the following steps: Suppose we want to test the following hypotheses:

$$H_0: \mu \leq \mu_0 \quad \text{versus} \quad H_a: \mu > \mu_0$$

1. Select a random sample $y_1, y_2, \ldots, y_n$ of size n from the population, and compute the value of $t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$.
2. Select a random sample of size n, with replacement from $y_1, y_2, \ldots, y_n$, and compute the mean $\bar{y}^*$ and standard deviation s^* of $y_1^*, y_2^*, \ldots, y_n^*$.
3. Compute the value of the statistic
$$\hat{t} = \frac{\bar{y}^* - \bar{y}}{s^*/\sqrt{n}}$$
4. Repeat Steps 2–4 a large number of times, B, to obtain $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_B$. Use these B values to approximate sampling distribution of $\frac{\bar{y} - \mu}{s/\sqrt{n}}$.
5. Let m be the number of values that are greater than or equal to the value t computed from the original sample.
6. The bootstrap p-value is $\frac{m}{B}$.

When the hypotheses are $H_0: \mu \geq \mu_0$ versus $H_a: \mu < \mu_0$, the only change would be to let m be the number of values from $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_B$ that are less than or equal to the value t computed from the original sample. Finally, when the hypotheses are $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$, let m_L be the number of values from $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_B$ that are less than or equal to the value t computed from the original sample and m_U be the number of values from $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_B$ that are greater than or equal to the value t computed from the original sample. Compute $p_L = \frac{m_L}{B}$ and $p_U = \frac{m_U}{B}$. Take the p-value to be the minimum of $2p_L$ and $2p_U$.

A point of clarification concerning the procedure described above: The bootstrap test statistic replaces μ_0 with the sample mean from the original sample. Recall that when we calculate the p-value of a test statistic, the calculation is always done under the assumption that the null hypothesis is true. In our bootstrap procedure, this requirement results in the bootstrap test statistic having μ_0 replaced with the sample mean from the original sample. This ensures that our bootstrap approximation of the sampling distribution of the test statistic is under the null value of μ, μ_0.

EXAMPLE 5.19

Refer to Example 5.18. CPS personnel wanted to determine if the mean cotanine level was greater than 75 for children under their supervision. Based on the sample of 20 children and using $\alpha = .05$, do the data support the contention that the mean exceeds 75?

Solution The set of hypotheses that we want to test is

$$H_0: \mu \leq 75 \quad \text{versus} \quad H_0: \mu > 75$$

Because there was a strong indication that the distribution of contanine levels in the population of children under CPS supervision was not normally distributed and because the sample size n was relatively small, the use of the t distribution to compute the p-value may result in a very erroneous decision based on the observed data. Therefore, we will use the bootstrap procedure.

First, we calculate the value of the test statistic in the original data:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{105.75 - 75}{82.429/\sqrt{20}} = 1.668$$

Next, we use the 9,999 bootstrap samples generated in Example 5.18 to determine the number of samples, m, with $\hat{t} = \frac{\bar{y}^* - \bar{y}}{s^*/\sqrt{n}} = \frac{\bar{y}^* - 105.75}{s^*/\sqrt{20}}$ greater than 1.668. From the 9,999 values of $\hat{t}$, we find that $m = 330$ of the $B = 9{,}999$ values of $\hat{t}$ exceeded or were equal to 1.668. Therefore, our p-value $= m/B = 330/9{,}999 = .033 < .05 = \alpha$. Therefore, we conclude that there is sufficient evidence that the mean cotanine level exceeds 75 in the population of children under CPS supervision.

It is interesting to note that if we had used the t distribution with 19 degrees of freedom to compute the p-value, the result would have produced a different conclusion. From Table 2 in the Appendix with df = 19,

$$p\text{-value} = P[t \geq 1.668] = .056 > .05 = \alpha$$

Using the t-tables, we would have concluded there is insufficient evidence in the data to support the contention that the mean cotanine exceeds 75. The small sample size, $n = 20$, and the possibility of nonnormal data would make this conclusion suspect. ■

Steps for Obtaining Bootstrap Tests and Confidence Intervals

The following steps using the R software will yield the p-value and confidence intervals given in Example 5.18 using $B = 9{,}999$ bootstrap samples selected with replacement from the original 20 data values. Note that each running of the code will yield slightly different values for the p-value and confidence intervals.

1. $x = c(29, 30, 53, 75, 89, 34, 21, 12, 58, 84, 92, 117, 115, 119, 109, 115, 134, 253, 289, 287)$
2. $n =$ **length**(x)
3. mndata = **mean**(x)
4. sdata = **sd**(x)
5. tdata = (mndata-75)/(sdata/**sqrt**(n))
6. $B = 9{,}999$

```
7. mnsamp = rep(0, times =B)
8. ssamp = rep(0, times = B)
9. tsamp = rep(0, times = B)
10. for (i in 1 : B) {
11. samp = sample(x, replace = TRUE)
12. mnsamp = mean(samp)
13. ssamp = sd(samp)
14. tsamp[i] = (mnsamp-mndata)/(ssamp/sqrt(n)) }
15. pval = sum(tsamp > = tdata)/B
16. tsort = sort(tsamp)
17. L = mndata − tsort[9750]*sdata/sqrt(n)
18. U = mndata − tsort[250]*sdata/sqrt(n)
```

5.9 Inferences About the Median

When the population distribution is highly skewed or very heavily tailed, the median is more appropriate than the mean as a representation of the center of the population. Furthermore, as was demonstrated in Section 5.7, the t procedures for constructing confidence intervals and for testing hypotheses for the population mean are not appropriate when applied to random samples from such populations with small sample sizes. In this section, we will develop a test of hypotheses and a confidence interval for the population median that will be appropriate for all types of population distributions.

The estimator of the population median M is based on the order statistics that were discussed in Chapter 3. Recall that if the measurements from a random sample of size n are given by $y_1, y_2, \ldots, y_n$, then the order statistics are these values ordered from smallest to largest. Let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ represent the data in ordered fashion. Thus, $y_{(1)}$ is the smallest data value and $y_{(n)}$ is the largest data value. The estimator of the population median is the sample median $\hat{M}$. Recall that $\hat{M}$ is computed as follows:

If n is an odd number, then $\hat{M} = y_{(m)}$, where $m = (n + 1)/2$.
If n is an even number, then $\hat{M} = (y_{(m)} + y_{(m+1)})/2$, where $m = n/2$.

To take into account the variability of $\hat{M}$ as an estimator of M, we next construct a confidence interval for M. A confidence interval for the population median M may be obtained by using the binomial distribution with $\pi = 0.5$.

100(1 − α)% Confidence Interval for the Median

A confidence interval for M with level of confidence at least $100(1 - \alpha)\%$ is given by

$$(M_L, M_U) = (y_{(L_{\alpha/2})}, y_{(U_{\alpha/2})})$$

where

$$L_{\alpha/2} = C_{\alpha(2),n} + 1$$
$$U_{\alpha/2} = n - C_{\alpha(2),n}$$

Table 4 in the Appendix contains values for $C_{\alpha(2),n}$, which are percentiles from a binomial distribution with $\pi = .5$.

Because the confidence limits are computed using the binomial distribution, which is a discrete distribution, the level of confidence of (M_L, M_U) will generally be somewhat larger than the specified $100(1 - \alpha)\%$. The exact level of confidence is given by

$$\text{Level} = 1 - 2P[\text{Bin}(n, .5) \leq C_{\alpha(2),n}] = 1 - 2\,\textbf{pbinom}(C_{\alpha(2),n}, n, .5)$$

The following example will demonstrate the construction of the interval.

EXAMPLE 5.20

The sanitation department of a large city wants to investigate ways to reduce the amount of recyclable materials that are placed in the city's landfill. By separating the recyclable material from the remaining garbage, the city could prolong the life of the landfill site. More important, the number of trees needed to be harvested for paper products and the aluminum needed for cans could be greatly reduced. From an analysis of recycling records from other cities, it is determined that if the average weekly amount of recyclable material is more than 5 pounds per household, a commercial recycling firm could make a profit collecting the material. To determine the feasibility of the recycling plan, a random sample of 25 households is selected. The weekly weight of recyclable material (in pounds/week) for each household is given here.

14.2	5.3	2.9	4.2	1.2	4.3	1.1	2.6	6.7	7.8	25.9	43.8	2.7
5.6	7.8	3.9	4.7	6.5	29.5	2.1	34.8	3.6	5.8	4.5	6.7	

Determine an appropriate measure of the amount of recyclable waste from a typical household in the city.

FIGURE 5.22(a)
Boxplot for waste data

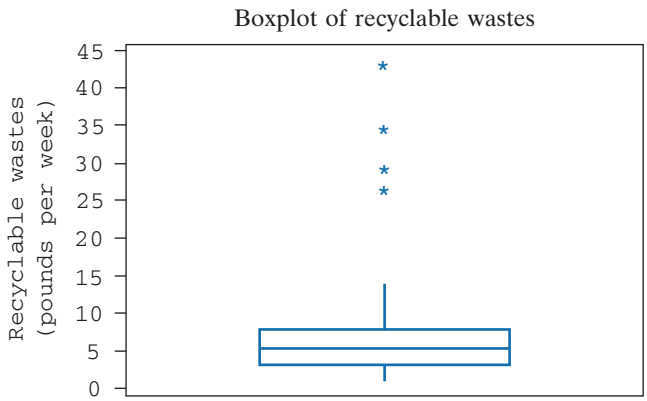
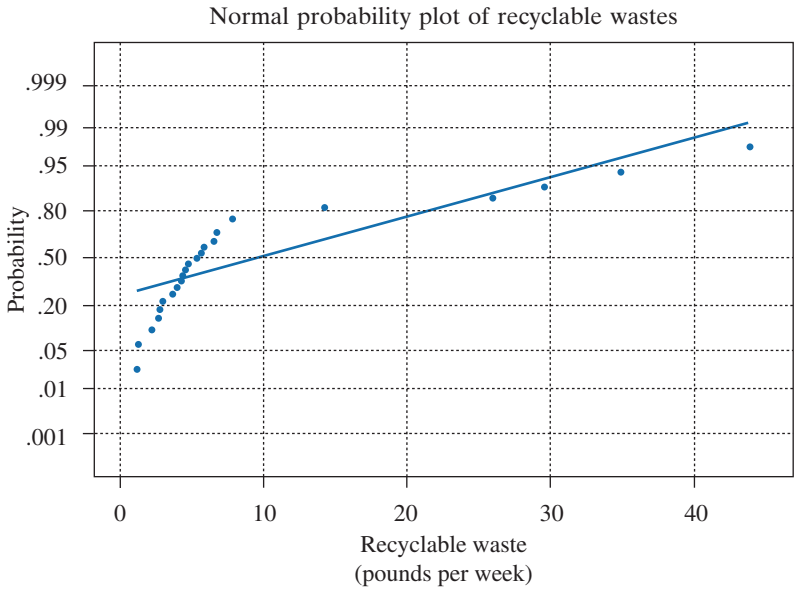


FIGURE 5.22(b)
Normal probability plot for waste data



Solution A boxplot and normal probability of the recyclable waste data (Figures 5.22(a) and (b)) reveal the extreme right skewness of the data. Thus, the mean is not an appropriate representation of the typical household's potential recyclable material. The sample median and a confidence interval on the population are given by the following computations. First, we order the data from smallest value to largest value:

1.1	1.2	2.1	2.6	2.7	2.9	3.6	3.9	4.2	4.3	4.5	4.7	5.3
5.6	5.8	6.5	6.7	6.7	7.8	7.8	14.2	25.9	29.5	34.8	43.8	

The number of values in the data set is an odd number, so the sample median is given by

$$\hat{M} = y_{((25+1)/2)} = y_{(13)} = 5.3$$

The sample mean is calculated to be $\bar{y} = 9.53$. Thus, we see that 20 of the 25 households have weekly recyclable waste that is less than the sample mean. Note that 12 of the 25 waste values are less and 12 of the 25 are greater than the sample median. Thus, the sample median is more representative of the typical household's recyclable waste than is the sample mean. Next, we will construct a 95% confidence interval for the population median.

From Table 4 in the Appendix, we find

$$C_{\alpha(2),\,n} = C_{.05,\,25} = 7$$

Thus,

$$L_{.025} = C_{.05,\,25} + 1 = 8$$
$$U_{.025} = n - C_{.05,\,n} = 25 - 7 = 18$$

The 95% confidence interval for the population median is given by

$$(M_L, M_U) = (y_{(8)}, y_{(18)}) = (3.9, 6.7)$$

Using the binomial distribution, the exact level of coverage is given by $1 - 2P[\text{Bin}(25, .5) \leq 7] = .957$, which is slightly larger than the desired level 95%. Thus, we are at least 95% confident that the median amount of recyclable waste per household is between 3.9 and 6.7 pounds per week. ■

Large-Sample Approximation

When the sample size n is large, we can apply the normal approximation to the binomial distribution to obtain approximations to $C_{\alpha(2),n}$. The approximate value is given by

$$C_{\alpha(2),\,n} \approx \frac{n}{2} - z_{\alpha/2}\sqrt{\frac{n}{4}}$$

Because this approximate value for $C_{\alpha(2),\,n}$ is generally not an integer, we set $C_{\alpha(2),\,n}$ to be the largest integer that is less than or equal to the approximate value.

EXAMPLE 5.21

Using the data in Example 5.20, find a 95% confidence interval for the median using the approximation to $C_{\alpha(2),\,n}$.

Solution We have $n = 25$ and $\alpha = .05$. Thus, $z_{.05/2} = 1.96$, and

$$C_{\alpha(2),\,n} \approx \frac{n}{2} - z_{\alpha/2}\sqrt{\frac{n}{4}} = \frac{25}{2} - 1.96\sqrt{\frac{25}{4}} = 7.6$$

Thus, we set $C_{\alpha(2),n} = 7$, and our confidence interval is identical to the interval constructed in Example 5.20. If n is larger than 30, the approximate and the exact value of $C_{\alpha(2),n}$ will often be the same integer. ■

In Example 5.20, the city wanted to determine whether the median amount of recyclable material was more than 5 pounds per household per week. We constructed a confidence interval for the median, but we still have not answered the question of whether the median is greater than 5. Thus, we need to develop a test of hypotheses for the median.

We will use the ideas developed for constructing a confidence interval for the median in our development of the testing procedures for hypotheses concerning a population median. In fact, a $100(1-\alpha)\%$ confidence interval for the population median M can be used to test two-sided hypotheses about M. If we want to test H_0: $M = M_0$ versus H_1: $M \neq M_0$ at level α, then we construct a $100(1-\alpha)\%$ confidence interval for M. If M_0 is contained in the confidence interval, then we fail to reject H_0. If M_0 is outside the confidence interval, then we reject H_0.

sign test

For testing one-sided hypotheses about M, we will use the binomial distribution to determine the rejection region. The testing procedure is called the **sign test** and is constructed as follows. Let $y_1, \ldots, y_n$ be a random sample from a population having median M. Let the null value of M be M_0, and define $W_i = y_i - M_0$. The sign test statistic B is the number of positive W_is. Note that B is simply the number of y_is that are greater than M_0. Because M is the population median, 50% of the data values are greater than M and 50% are less than M. Now, if $M = M_0$, then there is a 50% chance that y_i is greater than M_0 and hence a 50% chance that W_i is positive. Because the W_is are independent, each W_i has a 50% chance of being positive whenever $M = M_0$, and B counts the number of positive W_is under H_0. B is a binomial random variable with $\pi = .5$, and the percentiles from the binomial distribution with $\pi = .5$ given in Table 4 in the Appendix can be used to construct the rejection region for the test of hypotheses. The statistical **test for a population median *M*** is summarized next. Three different sets of hypotheses are given with their corresponding rejection regions. The tests given are appropriate for any population distribution.

test for a population median *M*

Summary of a Statistical Test for the Population Median *M*

Hypotheses:

Case 1. H_0: $M \leq M_0$ vs. H_a: $M > M_0$ (right-tailed test)
Case 2. H_0: $M \geq M_0$ vs. H_a: $M < M_0$ (left-tailed test)
Case 3. H_0: $M = M_0$ vs. H_a: $M \neq M_0$ (two-tailed test)

T.S.: Let $W_i = y_i - M_0$ and B = number of positive W_is.

R.R.: For a probability α of a Type I error,

Case 1. Reject H_0 if $B \geq n - C_{\alpha(1),n}$
Case 2. Reject H_0 if $B \leq C_{\alpha(1),n}$
Case 3. Reject H_0 if $B \leq C_{\alpha(2),n}$ or $B \geq n - C_{\alpha(2),n}$.

The following example will illustrate the test of hypotheses for the population median.

EXAMPLE 5.22

Refer to Example 5.20. The sanitation department wanted to determine whether the median household recyclable waste was greater than 5 pounds per week. Test this research hypothesis at level $\alpha = .05$ using the data from Exercise 5.20.

Solution The set of hypotheses is

$$H_0: M \leq 5 \text{ versus } H_a: M > 5$$

The data set consisted of a random sample of $n = 25$ households. From Table 4 in the Appendix, we find $C_{\alpha(1),n} = C_{.05,25} = 7$. Thus, we will reject $H_0: M \leq 5$ if $B \geq n - C_{\alpha(1),n} = 25 - 7 = 18$. Let $W_i = y_i - M_0 = y_i - 5$, which yields

−3.9	−3.8	−2.9	−2.4	−2.3	−2.1	−1.4	−1.1	−0.8
−0.7	−0.5	−0.3	0.3	0.6	0.8	1.5	1.7	1.7
2.8	2.8	9.2	20.9	24.5	29.8	38.8		

The 25 values of W_i contain 13 positive values. Thus, $B = 13$, which is not greater than 18. We conclude that the data set fails to demonstrate that the median household level of recyclable waste is greater than 5 pounds. ■

Large-Sample Approximation

When the sample size n is larger than the values given in Table 4 in the Appendix, we can use the normal approximation to the binomial distribution to set the rejection region. The standardized version of the sign test is given by

$$B_{ST} = \frac{B - (n/2)}{\sqrt{n/4}}$$

When M equals M_0, B_{ST} has approximately a standard normal distribution. Thus, we have the following decision rules for the three different research hypotheses:

Case 1. Reject $H_0: M \leq M_0$ if $B_{ST} \geq z_\alpha$, with p-value $= P(z \geq B_{ST})$

Case 2. Reject $H_0: M \geq M_0$ if $B_{ST} \leq -z_\alpha$, with p-value $= P(z \leq B_{ST})$

Case 3. Reject $H_0: M = M_0$ if $|B_{ST}| \geq z_{\alpha/2}$, with p-value $= 2P(z \geq |B_{ST}|)$

where z_α is the standard normal percentile.

EXAMPLE 5.23

Using the information in Example 5.22, construct the large-sample approximation to the sign test, and compare your results to those obtained using the exact sign test.

Solution Refer to Example 5.22, where we had $n = 25$ and $B = 13$. We conduct the large-sample approximation to the sign test as follows. We will reject $H_0: M \leq 5$ in favor of $H_\alpha: M > 5$ if $B_{ST} \geq z_{.05} = 1.96$.

$$B_{ST} = \frac{B - (n/2)}{\sqrt{n/4}} = \frac{13 - (25/2)}{\sqrt{25/4}} = 0.2$$

Because B_{ST} is not greater than 1.96, we fail to reject H_0. The p-value $= P(z \geq 0.2) = 1 - P(z < 0.2) = 1 - .5793 = .4207$ using Table 1 in the Appendix. Thus, we reach the same conclusion as was obtained using the exact sign test. ■

In Section 5.7, we observed that the performance of the t test deteriorated when the population distribution was either very heavily tailed or highly skewed. In Table 5.8, we compute the level and power of the sign test and compare these values to the comparable values for the t test for the four population distributions depicted in Figure 5.19 in Section 5.7. Ideally, the level of the test should remain the same for all population distributions. Also, we want tests having the largest

TABLE 5.8 Level and power values of the t test versus the sign test

Population Distribution	Test Statistic	$n = 10$ $(M_a - M_0)/\sigma$				$n = 15$ $(M_a - M_0)/\sigma$				$n = 20$ $(M_a - M_0)/\sigma$			
		Level	.2	.6	.8	Level	.2	.6	.8	Level	.2	.6	.8
Normal	t	.05	.145	.543	.754	.05	.182	.714	.903	.05	.217	.827	.964
	Sign	.055	.136	.454	.642	.059	.172	.604	.804	.058	.194	.704	.889
Heavily Tailed	t	.035	.104	.371	.510	.049	.115	.456	.648	.045	.163	.554	.736
	Sign	.055	.209	.715	.869	.059	.278	.866	.964	.058	.325	.935	.990
Lightly Skewed	t	.055	.140	.454	.631	.059	.178	.604	.794	.058	.201	.704	.881
	Sign	.025	.079	.437	.672	.037	.129	.614	.864	.041	.159	.762	.935
Highly Skewed	t	.007	.055	.277	.463	.006	.078	.515	.733	.011	.104	.658	.873
	Sign	.055	.196	.613	.778	.059	.258	.777	.912	.058	.301	.867	.964

possible power values because the power of a test is its ability to detect false null hypotheses. When the population distribution is either heavily tailed or highly skewed, the level of the t test changes from its stated value of .05. In these situations, the level of the sign test stays the same because the level of the sign test is the same for all distributions. The power of the t test is greater than the power of the sign test when sampling from a population having a normal distribution. However, the power of the sign test is greater than the power of the t test when sampling from very heavily tailed distributions or highly skewed distributions.

5.10 RESEARCH STUDY: Percentage of Calories from Fat

In Section 5.1, we introduced the potential health problems associated with obesity. The assessment and quantification of a person's usual diet is crucial in evaluating the degree of relationship between diet and diseases. This is a very difficult task but is important in an effort to monitor dietary behavior among individuals. ***Rosner, Willett, and Spiegelman, in "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Systematic Within-Person Measurement Error" [Statistics in Medicine (1989) 8:1051–1070]***, describe a nurses' health study in which the diet of a large sample of women was examined. One of the objectives of the study was to determine the percentage of calories from fat (PCF) in the diet of a population of nurses and compare this value with the recommended value of 30%. The most commonly used method in large nutritional epidemiology studies is the food frequency questionnaire (FFQ). This questionnaire uses a carefully designed series of questions to determine the dietary intakes of participants in the study. In the nurses' health study, a sample of nurses completed a single FFQ. These women represented a random sample from a population of nurses. From the information gathered from the questionnaire, the PCF was then computed.

To minimize missteps in a research study, it is advisable to follow the four-step process outlined in Chapter 1. We will illustrate these steps using the PCF study described at the beginning of this chapter. The first step is determining the goals and objectives of the study.

Defining the Problem

The researchers in this study would need to answer questions similar to the following:

1. What is the population of interest?
2. What dietary variables may have an effect on a person's health?

3. What characteristics of the nurses other than dietary intake may be important in studying their health condition?
4. How should the nurses be selected to participate in the study?
5. What hypotheses are of interest to the researchers?

The researchers decided that the main variable of interest was the percentage of calories from fat in the diet of nurses. The parameters of interest were the PCF mean μ for the population of nurses, the standard deviation σ of the PCF for the population of nurses, and the proportion π of nurses having a PCF greater than 50%. They also wanted to determine if the average PCF for the population of nurses exceeded the recommended value of 30%.

In order to estimate these parameters and test hypotheses about the parameters, it was first necessary to determine the sample size required to meet certain specifications imposed by the researchers. The researchers wanted to estimate the mean PCF with a 95% confidence interval having a tolerable error of 3. From previous studies, the PCF values ranged from 10% to 50%. Because we want a 95% confidence interval with width 3, $E = 3/2 = 1.5$ and $z_{\alpha/2} = z_{.025} = 1.96$. Our estimate of σ is $\hat{\sigma} = \text{range}/4 = (50 - 10)/4 = 10$. Substituting into the formula for n, we have

$$n = \frac{(z_{\alpha/2})^2 \hat{\sigma}^2}{E^2} = \frac{(1.96)^2(10)^2}{(1.5)^2} = 170.7$$

Thus, a random sample of 171 nurses should give a 95% confidence interval for μ with the desired width of 3, provided 10 is a reasonable estimate of σ. Three nurses originally selected for the study did not provide information on PCF; therefore, the sample size was only 168.

Collecting the Data

The researchers would need to carefully examine the data from the FFQs to determine if the responses were recorded correctly. The data would then be transfered to computer files and prepared for analysis following the steps outlined in Chapter 2. The next step in the study would be to summarize the data through plots and summary statistics.

Summarizing the Data

The PCF values for the 168 women are displayed in Figure 5.23 in a stem-and-leaf diagram along with a table of summary statistics. A normal probability plot is provided in Figure 5.24 to assess the normality of the distribution of PCF values.

From the stem-and-leaf plot and normal probability plot, it appears that the data are nearly normally distributed, with PCF values ranging from 15% to 57%. The proportion of the women who have a PCF greater than 50% is $\hat{\pi} = 4/168 = 2.4$ %. From the table of summary statistics in the output, the sample mean is $\bar{y} = 36.919$, and the sample standard deviation is $s = 6.728$. The researchers want to draw inferences from the random sample of 168 women to the population from which they were selected. Thus, we would need to place bounds on our point estimates in order to reflect our degree of confidence in their estimation of the population values. Also, the researchers may be interested in testing hypotheses about the size of the population PCF mean μ or variance σ^2. For example, many nutritional experts recommend that one's daily diet have no more than 30% of total calories from fat. Thus, we would want to test the statistical hypothesis that μ is greater than 30 to determine if the average PCF value for the population of nurses exceeds the recommended value.

FIGURE 5.23 The percentage of calories from fat (PCF) for 168 women in a dietary study

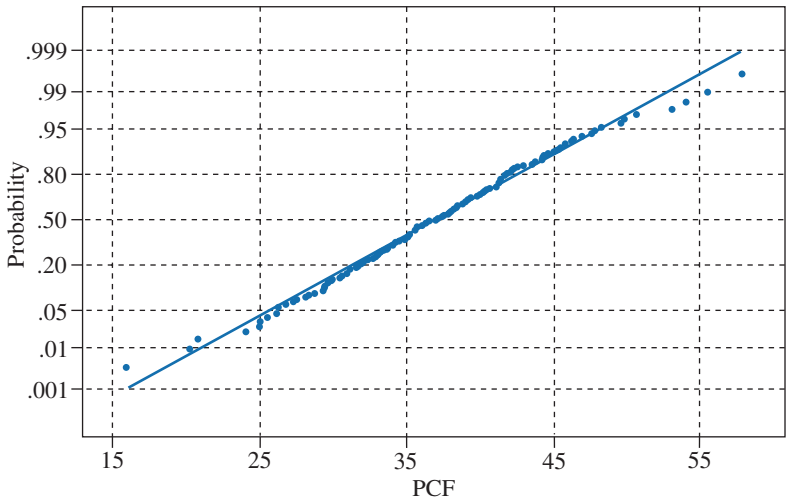
```
1   5
2   0 0 4 4
2   5 5 6 6 6 6 7 7 8 8 8 9 9 9 9 9 9 9 9
3   0 0 0 0 0 0 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4
3   5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9 9
4   0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 4 4 4 4 4
4   5 5 5 5 5 6 6 6 6 7 7 8 9 9
5   0 3 4
5   5 7
```

Descriptive Statistics for Percentage of Calories from Fat Data

Variable	N	Mean	Median	TrMean	StDev	SE Mean
PCF	168	36.919	36.473	36.847	6.728	0.519

Variable	Minimum	Maximum	Q1	Q3
PCF	15.925	57.847	32.766	41.295

FIGURE 5.24
Normal probability plot for percentage of calories from fat (PCF)



Analyzing the Data and Interpreting the Analyses

One of the objectives of the study was to estimate the mean PCF in the diet of nurses. Also, the researchers wanted to test whether the mean was greater than the recommended value of 30%. Prior to constructing confidence intervals or testing hypotheses, we must first check whether the data represent a random sample from a normally distributed population. From the normal probability plot in Figure 5.24, the data values fall nearly on a straight line. Hence, we can conclude that the data appear to follow a normal distribution. The mean and standard deviation of the PCF data were given by $\bar{y} = 36.92$ and $s = 6.73$. We can next construct a 95% confidence interval for the mean PCF for the population of nurses as follows:

$$36.92 \pm t_{.025,\,167}\frac{6.73}{\sqrt{168}} = 36.92 \pm 1.974\,\frac{6.73}{\sqrt{168}} = 36.92 \pm 1.02$$

Thus, we are 95% confident that the mean PCF in the population of nurses is between 35.90 and 37.94. As a result, we would be inclined to conclude that the mean PCF for the population of nurses exceeds the recommended value of 30.

We will next formally test the following hypotheses:

$$H_0: \mu \leq 30 \quad \text{versus} \quad H_a: \mu > 30$$

Since the data appear to be normally distributed and in any case the sample size is reasonably large, we can use the t test with rejection region as follows.

R.R: For a one-tail t test with $\alpha = .05$, we reject H_0 if

$$t = \frac{\bar{y} - 30}{s/\sqrt{168}} \geq t_{.05,\,167} = 1.654$$

Since $t = \frac{36.92 - 30}{6.73/\sqrt{168}} = 13.33$, we reject H_0. The p-value of the test is essentially 0, so we can conclude that the mean PCF value is very significantly greater than 30. Thus, there is strong evidence that the population of nurses has an average PCF larger than the recommended value of 30. The experts in this field would have to determine the practical consequences of having a PCF value between 5.90 and 7.94 units higher than the recommended value.

Reporting the Conclusions

A report summarizing our findings from the study would include the following items:

1. Statement of objective for study
2. Description of study design and data collection procedures
3. Numerical and graphical summaries of data sets
4. Description of all inference methodologies:
   - t tests
   - t-based confidence interval on population mean
   - Verification that all necessary conditions for using inference techniques were satisfied
5. Discussion of results and conclusions
6. Interpretation of findings relative to previous studies
7. Recommendations for future studies
8. Listing of data set

5.11 Summary and Key Formulas

A population mean or median can be estimated using point or interval estimation. The selection of the median in place of the mean as a representation of the center of a population depends on the shape of the population distribution. The performance of an interval estimate is determined by the width of the interval and the confidence coefficient. The formulas for a $100(1 - \alpha)\%$ confidence interval for the mean μ and median M were given. A formula was provided for determining the necessary sample size in a study so that a confidence interval for μ would have a predetermined width and level of confidence.

Following the traditional approach to hypothesis testing, a statistical test consists of five parts: research hypothesis, null hypothesis, test statistic, rejection region, and checking assumptions and drawing conclusions. A statistical test employs the technique of proof by contradiction. We conduct experiments and studies to gather data to verify the research hypothesis through the contradiction of the null hypothesis H_0. As with any two-decision process based on variable data, there are two types of errors that can be committed. A Type I error is the rejection of H_0 when H_0 is true, and a Type II error is the acceptance of H_0 when the alternative hypothesis H_a is true.

The probability for a Type I error is denoted by α. For a given value of the mean μ_a in H_a, the probability of a Type II error is denoted by $\beta(\mu_a)$. The value of $\beta(\mu_a)$ decreases as the distance from μ_a to μ_0 increases. The power of a test of hypotheses is the probability that the test will reject H_0 when the value of μ resides in H_a. Thus, the power at μ_a equals $1 - \beta(\mu_a)$.

We also demonstrated that for a given sample size and value of the mean μ_a, α and $\beta(\mu_a)$ are inversely related; as α is increased, $\beta(\mu_a)$ decreases, and vice versa. If we specify the sample size n and α for a given test procedure, we can compute $\beta(\mu_a)$ for values of the mean μ_a in the alternative hypothesis. In many studies, we need to determine the necessary sample size n to achieve a testing procedure having a specified value for α and a bound on $\beta(\mu_a)$. A formula is provided to determine n such that a level α test has $\beta(\mu_a) \leq \beta$ whenever μ_a is a specified distance beyond μ_0.

We developed an alternative to the traditional decision-based approach for a statistical test of hypotheses. Rather than relying on a preset level of α, we compute the weight of evidence in the data for rejecting the null hypothesis. This weight, expressed in terms of a probability, is called the level of significance for the test. Most professional journals summarize the results of a statistical test using the level of significance. We discussed how the level of significance can be used to obtain the same results as the traditional approach.

We also considered inferences about μ when σ is unknown (which is the usual situation). Through the use of the t distribution, we can construct both confidence intervals and a statistical test for μ. The t-based tests and confidence intervals do not have the stated levels or power when the population distribution is highly skewed or very heavily tailed and the sample size is small. In these situations, we may use the median in place of the mean to represent the center of the population. Procedures were provided to construct confidence intervals and tests of hypotheses for the population median. Alternatively, we can use bootstrap methods to approximate confidence intervals and tests when the population distribution is nonnormal and n is small.

Key Formulas

Estimation and tests for μ and the median:

1. $100(1 - \alpha)\%$ confidence interval for μ (σ unknown) when sampling from a normal population or when n is large

$$\bar{y} \pm t_{\alpha/2} s/\sqrt{n}, \text{ df} = n - 1$$

2. Sample size for estimating μ with a $100(1 - \alpha)\%$ confidence interval, $\bar{y} \pm E$

$$n = \frac{(z_{\alpha/2})^2 \hat{\sigma}^2}{E^2}$$

where $\hat{\sigma}^2$ is an estimate of population variance.

3. Statistical test for μ (σ unknown) when sampling from a normal population or when n is large

$$\text{Test statistics: } t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}, \text{ df} = n - 1$$

4. Calculation of $\beta(\mu_a)$ (and equivalent power) for a test on μ ($\hat{\sigma}$ estimate of σ) when sampling from a normal population or when n is large
   a. One-tailed level α test

$$\beta(\mu_a) = P\left(z < z_\alpha - \frac{|\mu_0 - \mu_a|}{\hat{\sigma}/\sqrt{n}}\right)$$

b. Two-tailed level α test

$$\beta(\mu_a) \approx P\left(z < z_{\alpha/2} - \frac{|\mu_0 - \mu_a|}{\hat{\sigma}/\sqrt{n}}\right)$$

5. Calculation of $\beta(\mu_a)$ (and equivalent power) for a test on μ (σ unknown) when sampling from a normal population or when n is large: Use Table 3 in the Appendix.

6. Sample size n for a statistical test on μ ($\hat{\sigma}$ estimate of σ) when sampling from a normal population

a. One-tailed level α test

$$n = \frac{\hat{\sigma}^2}{\Delta^2}(z_\alpha + z_\beta)^2$$

b. Two-tailed level α test

$$n > \frac{\hat{\sigma}^2}{\Delta^2}(z_{\alpha/2} + z_\beta)^2$$

7. $100(1 - \alpha)\%$ confidence interval for the population median M

$$(y_{(L_{\alpha/2})}, y_{(U_{\alpha/2})}), \text{ where } L_{\alpha/2} = C_{\alpha(2),\,n} + 1 \text{ and } U_{\alpha/2} = n - C_{\alpha(2),\,n}$$

8. Statistical test for median

Test statistic:

Let $W_i = y_i - M_0$ and B = number of positive W_is

5.12 Exercises

5.1 Introduction

Pol. Sci. **5.1** The county government in a city that is dominated by a large state university is concerned that a small subset of its population has been overutilized in the selection of residents to serve on county court juries. The county decides to determine the mean number of times that an adult resident of the county has been selected for jury duty during the past 5 years. They will then compare the mean jury participation for full-time students to that of nonstudents.

a. Identify the populations of interest to the county officials.

b. How might you select a sample of voters to gather this information?

Med. **5.2** In the research study on percentage of calories from fat,

a. What is the population of interest?

b. What dietary variables other than PCF might affect a person's health?

c. What characteristics of the nurses other than dietary intake might be important in studying their health condition?

d. Describe a method for randomly selecting which nurses participate in the study.

e. State several hypotheses that may be of interest to the researchers.

Engin. **5.3** Face masks used by firefighters often fail by having their lenses fall out when exposed to very high temperatures. A manufacturer of face masks claims that for its masks the average temperature at which pop-out occurs is 550°F. A sample of 75 masks is tested, and the average temperature at which the lenses popped out was 470°F. Based on this information is the manufacturer's claim valid?

a. Identify the population of interest to the firefighters in this problem.

b. Would an answer to the question posed involve estimation or hypothesis testing?

5.4 Refer to Exercise 5.3. Describe a process to select a sample of face masks from the manufacturer to evaluate the claim.

5.2 Estimation of μ

Engin. **5.5** A company that manufacturers coffee for use in commercial machines monitors the caffeine content in its coffee. The company selects 50 samples of coffee every hour from its production line and determines the caffeine content. From historical data, the caffeine content (in milligrams, mg) is known to have a normal distribution with $\hat{\sigma} = 7.1$ mg. During a 1-hour time period, the 50 samples yielded a mean caffeine content of $\bar{y} = 110$ mg.

- **a.** Identify the population about which inferences can be made from the sample data.
- **b.** Calculate a 95% confidence interval for the mean caffeine content μ of the coffee produced during the hour in which the 50 samples were selected.
- **c.** Explain to the CEO of the company in nonstatistical language the interpretation of the constructed confidence interval.

5.6 Refer to Exercise 5.5. The engineer in charge of the coffee manufacturing process examines the confidence intervals for the mean caffeine content calculated over the past several weeks and is concerned that the intervals are too wide to be of any practical use. That is, they are not providing a very precise estimate of μ.

- **a.** What would happen to the width of the confidence intervals if the level of confidence of each interval is increased from 95% to 99%?
- **b.** What would happen to the width of the confidence intervals if the number of samples per hour was increased from 50 to 100?

5.7 Refer to Exercise 5.5. Because the company is sampling the coffee production process every hour, there are 720 confidence intervals for the mean caffeine content μ constructed every month.

- **a.** If the level of confidence remains at 95% for the 720 confidence intervals in a given month, how many of the confidence intervals would you expect to fail to contain the value of μ and hence provide an incorrect estimation of the mean caffeine content?
- **b.** If the number of samples is increased from 50 to 100 each hour, how many of the 95% confidence intervals would you expect to fail to contain the value of μ in a given month?
- **c.** If the number of samples remains at 50 each hour but the level of confidence is increased from 95% to 99% for each of the intervals, how many of the 99% confidence intervals would you expect to fail to contain the value of μ in a given month?

Bus. **5.8** As part of the recruitment of new businesses, the city's economic development department wants to estimate the gross profit margin of small businesses (under $1 million in sales) currently residing in the city. A random sample of the previous years annual reports of 15 small businesses shows the mean net profit margin to be 7.2% (of sales) with a standard deviation of 12.5%.

- **a.** Construct a 99% confidence interval for the mean gross profit margin of μ of all small businesses in the city.
- **b.** The city manager reads the report and states that the confidence interval for μ constructed in part (a) is not valid because the data are obviously not normally distributed and thus the sample size is too small. Based on just knowing the mean and standard deviation of the sample of 15 businesses, do you think the city manager is valid in his conclusion about the data? Explain your answer.

Soc. **5.9** A program to reduce recidivism has been in effect for two years in a large northeastern state. A sociologist investigates the effectiveness of the program by taking a random sample of 200 prison records of repeat offenders. The records were selected from the files in the courthouse of the largest city in the state. The average length of time out of prison between the first and second offenses is 2.8 years with a standard deviation of 1.3 years.

- **a.** Use this information to estimate the mean prison-free time between first and second offenses using a 95% confidence interval.

b. Identify the group for which the confidence interval would be an appropriate estimate of the population mean.
c. Would it be valid to use this confidence interval to estimate the mean prison-free time between first and second offenses for all two-time offenders in the whole state? In a large southern state?

Ag. **5.10** The susceptibility of the root stocks of a variety of orange tree to a specific larva is investigated by a group of researchers. Forty orange trees are exposed to the larva and then examined by the researchers 6 months after exposure. The number of larvae per gram is recorded on each root stock. The mean and standard deviation of the logarithm of the counts are recorded to be 9.02 and 1.12, respectively.

a. Use the sample information to construct a 90% confidence interval on the mean of the logarithm of the larvae counts.
b. Identify the population for which this confidence interval could be used to assess the susceptibility of the orange trees to the larva.

5.11 Refer to Example 5.4. Suppose an estimate of σ is given by $\hat{\sigma} = .7$.

a. If the level of confidence remains 99% but the desired width of the interval is reduced to 0.3, what is the necessary sample size?
b. If the level of confidence is reduced to 95% but the desired width of the interval remains 0.5, what is the necessary sample size?
c. If the level of confidence is increased to 99.5% but the desired width of the interval remains 0.5, what is the necessary sample size?
d. Describe the impact on the value of the sample size of increases (decreases) in the level of confidence for a fixed desired width.
e. Describe the impact on the value of the sample size of increases (decreases) in the desired width for a fixed level of confidence.

5.3 Choosing the Sample Size for Estimating μ

5.12 In any given situation, if the level of confidence and the standard deviation are kept constant, how much would you need to increase the sample size to decrease the width of the interval to half its original size?

Bio. **5.13** A biologist wishes to estimate the effect of an antibiotic on the growth of a particular bacterium by examining the mean amount of bacteria present per plate of culture when a fixed amount of the antibiotic is applied. Previous experimentation with the antibiotic on this type of bacteria indicates that the standard deviation of the amount of bacteria present is approximately 13 cm^2. Use this information to determine the number of observations (cultures that must be developed and then tested) necessary to estimate the mean amount of bacteria present, using a 99% confidence interval with a half-width of 3 cm^2.

Gov. **5.14** The housing department in a large city monitors the rent for rent-controlled apartments in the city. The mayor wants an estimate of the average rent. The housing department must determine the number of apartments to include in a survey in order to be able to estimate the average rent to within \$100 using a 95% confidence interval. From past surveys, the monthly charge for rent-controlled apartments ranged from \$1,000 to \$3,500. How many renters must be included in the survey to meet the requirements?

Gov. **5.15** Refer to Exercise 5.14. Suppose the mayor's staff reviews the proposed survey and decides that in order for the survey to be taken seriously the requirements need to be increased.

a. If the level of confidence is increased to 99% with the average rent estimated within \$50, how many apartments need to be included in the survey?
b. Suppose the budget for the survey will not support increasing the level of confidence to 99%. Provide an explanation to the mayor, who has never taken a statistics course, of the impact on the accuracy of the estimate of the average rent of not raising the level of confidence from 95% to 99%.

5.4 A Statistical Test for μ

Basic **5.16** A study is designed to test the hypotheses H_0: $\mu \geq 26$ versus H_a: $\mu < 26$. A random sample of 50 units was selected from a specified population, and the measurements were summarized to $\bar{y} = 25.9$ and $s = 7.6$.

- **a.** With $\alpha = .05$, is there substantial evidence that the population mean is less than 26?
- **b.** Calculate the probability of making a Type II error if the actual value of the population mean is at most 24.
- **c.** If the sample size is doubled to 100, what is the probability of making a Type II error if the actual value of the population mean is at most 24?

Basic **5.17** Refer to Exercise 5.16. Graph the power curve for rejecting H_0: $\mu \geq 26$ for the following values of μ: 20, 21, 22, 23, 24, 25, and 26.

- **a.** Describe the change in the power as the value of μ decreases from $\mu_0 = 26$.
- **b.** Suppose the value of n remains at 50 but α is decreased to $\alpha = .01$. Without recalculating the values of the power, superimpose on the graph for $\alpha = .05$ and $n = 50$ the power curve for $\alpha = .01$ and $n = 50$.
- **c.** Suppose the value of n is decreased to 35 but α is kept at $\alpha = .05$. Without recalculating the values of the power, superimpose on the graph for $\alpha = .05$ and $n = 50$ the power curve for $\alpha = .05$ and $n = 35$.

Basic **5.18** Use a computer to simulate 100 samples of $n = 25$ from a normal distribution with $\mu = 43$ and $\alpha = 4$. Test the hypotheses H_0: $\mu = 43$ versus H_a: $\mu \neq 43$ separately for each of the 100 samples of size 25 with $\alpha = .05$.

- **a.** How many of the 100 tests of hypotheses resulted in a rejection of H_0?
- **b.** Suppose 1,000 tests of hypotheses of H_0: $\mu = 43$ versus H_a: $\mu \neq 43$ were conducted. Each of the 1,000 data sets consists of $n = 50$ data values randomly selected from a population having $\mu = 43$. Suppose $\alpha = .05$ is used in each of the 1,000 tests. On the average, how many of the 1,000 tests would result in the rejection of H_0?
- **c.** Suppose the procedure in part (b) is repeated with 1,000 tests with $n = 75$ and $\alpha = .01$. On the average, how many of the 1,000 tests would result in a rejection of H_0?

Basic **5.19** Refer to Exercise 5.18. Simulate 100 samples of size $n = 25$ from a normal population in which $\mu = 45$ and $\sigma = 4$. Use $\alpha = .05$ in conducting a test of H_0: $\mu = 43$ versus H_a: $\mu \neq 43$ for each of the 100 samples.

- **a.** What proportion of the 100 tests of H_0: $\mu = 43$ versus H_a: $\mu \neq 43$ resulted in the correct decision, that is, the rejection of H_0?
- **b.** Calculate the power of the test of hypotheses necessary to reject H_0: $\mu = 43$ when the value of μ is 45.
- **c.** Based on the calculated probability, in part (b), how many of the 100 tests on the average should produce a rejection of H_0? Compare this value to the number of rejections obtained in the simulation. Explain why the estimated number of rejections and the number of rejections observed in the simulation differ.

Basic **5.20** Refer to Exercises 5.18 and 5.19.

- **a.** Answer the questions asked in Exercises 5.18 and 5.19 with $\alpha = .01$ replacing $\alpha = .05$. You can use the same simulated data, but the exact power will need to be recalculated.
- **b.** Did decreasing α from .05 to .01 result in the power increasing or decreasing? Explain why this change occurred.

Med. **5.21** A study was conducted of 90 adult male patients following a new treatment for congestive heart failure. One of the variables measured on the patients was the increase in exercise capacity (in minutes) over a 4-week treatment period. The previous treatment regime had produced an average increase of $\mu = 2$ minutes. The researchers wanted to evaluate whether the new

treatment had increased the value of μ in comparison to the previous treatment. The data yielded $\bar{y} = 2.17$ and $s = 1.05$.

a. Using $\alpha = .05$, what conclusions can you draw about the research hypothesis?

b. What is the probability of making a Type II error if the actual value of μ is 2.1?

5.22 Refer to Exercise 5.21. Compute the power of the test PWR(μ_a) at $\mu_a = 2.1, 2.2, 2.3, 2.4$, and 2.5. Sketch a smooth curve through a plot of PWR(μ_a) versus μ_a.

a. If α is reduced from .05 to .01, what would be the effect on the power curve?

b. If the sample size is reduced from 90 to 50, what would be the effect on the power curve?

5.5 Choosing the Sample Size for Testing μ

Med.

5.23 A national agency sets recommended daily dietary allowances for many supplements. In particular, the allowance for zinc for males over the age of 50 years is 15 mg/day. The agency would like to determine if the dietary intake of zinc for active males is significantly higher than 15 mg/day. How many males would need to be included in the study if the agency wants to construct an $\alpha = .05$ test with the probability of committing a Type II error at most .10 whenever the average zinc content is 15.3 mg/day or higher? Suppose from previous studies they estimate the standard deviation to be approximately 4 mg/day.

Edu.

5.24 To evaluate the success of a 1-year experimental program designed to increase the mathematical achievement of underprivileged high school seniors, a random sample of participants in the program will be selected and their mathematics scores will be compared with the previous year's statewide average of 525 for underprivileged seniors. The researchers want to determine whether the experimental program has increased the mean achievement level over the previous year's statewide average. If $\alpha = .05$, what sample size is needed to have a probability of Type II error of at most .025 if the actual mean is increased to 550? From previous results, $\sigma \approx 80$.

5.25 Refer to Exercise 5.24. Suppose a random sample of 100 students is selected yielding $\bar{y} = 542$ and $s = 76$. Is there sufficient evidence to conclude that the mean mathematics achievement level has been increased? Explain.

Bus.

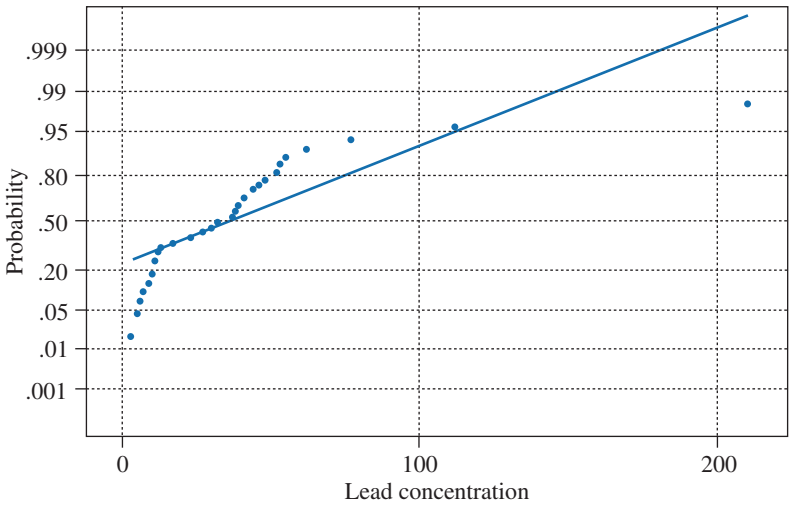
5.26 The administrator of a nursing home would like to do a time-and-motion study of staff time spent per day performing nonemergency tasks. Prior to the introduction of some efficiency measures, the average number of person-hours per day spent on these tasks was $\mu = 16$. The administrator wants to test whether the efficiency measures have reduced the value of μ. How many days must be sampled to test the proposed hypothesis if she wants a test having $\alpha = .05$ and the probability of a Type II error of at most .10 when the actual value of μ is 12 hours or less (at least a 25% decrease from the number of hours spent before the efficiency measures were implemented)? Assume $\sigma = 7.64$.

Env.

5.27 The vulnerability of inshore environments to contamination due to urban and industrial expansion in Mombasa is discussed in the paper ***"Metals, Petroleum Hydrocarbons and Organochlorines in Inshore Sediments and Waters on Mombasa, Kenya" [Marine Pollution Bulletin (1997) 34:570–577]***. A geochemical and oceanographic survey of the inshore waters of Mombasa, Kenya, was undertaken during the period from September 1995 to January 1996. In the survey, suspended particulate matter and sediment were collected from 48 stations within Mombasa's estuarine creeks. The concentrations of major oxides and 13 trace elements were determined for a varying number of cores at each of the stations. In particular, the lead concentrations in suspended particulate matter (mg kg^{-1} dry weight) were determined at 37 stations. The researchers were interested in determining whether the average lead concentration was greater than 30 mg kg^{-1} dry weight. The data are given in the following table along with summary statistics and a normal probability plot.

Lead concentrations (mg kg^{-1} dry weight) from 37 stations in Kenya

48	53	44	55	52	39	62	38	23	27
41	37	41	46	32	17	32	41	23	12
3	13	10	11	5	30	11	9	7	11
77	210	38	112	52	10	6			



a. Is there sufficient evidence ($\alpha = .05$) in the data that the mean lead concentration exceeds 30 mg kg^{-1} dry weight?
b. What is the probability of a Type II error if the actual mean concentration is 50?
c. Do the data appear to have a normal distribution?
d. Based on your answer in (c), is the sample size large enough for the test procedures to be valid? Explain.

5.6 The Level of Significance of a Statistical Test

Engin. **5.28** The R&D department of a paint company has developed an additive that it hopes will increase the ability of the company's stain for outdoor decks to resist water absorption. The current formulation of the stain has a mean absorption rate of 35 units. Before changing the stain, a study was designed to evaluate whether the mean absorption rate of the stain with the additive was decreased from the current rate of 35 units. The stain with the additive was applied to 50 pieces of decking material. The resulting data were summarized to $\bar{y} = 33.6$ and $s = 9.2$

a. Is there substantial evidence ($\alpha = .01$) that the additive reduces the mean absorption from its current value?
b. What is the level of significance (p-value) of your test results?
c. What is the probability of a Type II error if the stain with the additive in fact has a mean absorption rate of 30?
d. Estimate the mean absorption using a 99% confidence interval. Is the confidence interval consistent with your conclusions from the test of hypotheses?

Engin. **5.29** Refer to Exercise 5.28. If the R&D department used $\alpha = .10$ in place of $\alpha = .01$, would the conclusion about whether the additive reduced the mean absorption change from the conclusion using $\alpha = .01$?

Env. **5.30** A concern to public health officials is whether a concentration of lead in the paint of older homes may have an effect on the muscular development of young children. In order to evaluate this phenomenon, a researcher exposed 90 newly born mice to paint containing a specified amount of lead. The number of Type 2 fibers in the skeletal muscle was determined 6 weeks after exposure. The mean number of Type 2 fibers in the skeletal muscles of normal mice of this age is 21.7. The $n = 90$ mice yielded $\bar{y} = 18.8$, $s = 15.3$. Is there significant evidence in the data to support the hypothesis that the mean number of Type 2 fibers is different from 21.7 using an $\alpha = .05$ test?

5.31 Refer to Exercise 5.30. In fact, the researcher was more concerned about determining if the lead in the paint reduced the mean number of Type 2 fibers in skeletal muscles. Does the change in the research hypothesis alter your conclusion about the effect of lead in paint on the mean number of Type 2 fibers in skeletal muscles?

Med. **5.32** A tobacco company advertises that the average nicotine content of its cigarettes is at most 14 milligrams. A consumer protection agency wants to determine whether the average nicotine content is in fact greater than 14. A random sample of 300 cigarettes of the company's brand yields an average nicotine content of 14.6 milligrams and a standard deviation of 3.8 milligrams. Determine the level of significance of the statistical test of the agency's claim that μ is greater than 14. If $\alpha = .01$, is there significant evidence that the agency's claim has been supported by the data?

Psy. **5.33** A psychological experiment was conducted to investigate the length of time (time delay) between the administration of a stimulus and the observation of a specified reaction. A random sample of 36 persons was subjected to the stimulus, and the time delay was recorded. The sample mean and standard deviation were 2.2 and .57 seconds, respectively. Is there significant evidence that the mean time delay for the hypothetical population of all persons who may be subjected to the stimulus differs from 1.6 seconds? Use $\alpha = .05$. What is the level of significance of the test?

5.7 Inferences About μ for a Normal Population, σ Unknown

Basic **5.34** Provide the rejection region based on a t-test statistic for the following situations:

a. H_0: $\mu \geq 28$ versus H_a: $\mu < 28$ with $n = 11, \alpha = .05$
b. H_0: $\mu \leq 28$ versus H_a: $\mu > 28$ with $n = 21, \alpha = .025$
c. H_0: $\mu \geq 28$ versus H_a: $\mu < 28$ with $n = 8, \alpha = .001$
d. H_0: $\mu = 28$ versus H_a: $\mu \neq 28$ with $n = 13, \alpha = .01$

Basic **5.35** A study was designed to evaluate whether the population of interest has a mean greater than 9. A random sample of $n = 17$ units was selected from a population, and the data yield $\bar{x} = 10.1$ and $s = 3.1$.

a. Is there substantial evidence ($\alpha = .05$) that the population mean is greater than 9?
b. What is the level of significance of the test?

Edu. **5.36** The ability to read rapidly and simultaneously maintain a high level of comprehension is often a determining factor in the academic success of many high school students. A school district is considering a supplemental reading program for incoming freshmen. Prior to implementing the program, the school runs a pilot program on a random sample of $n = 20$ students. The students were thoroughly tested to determine reading speed and reading comprehension. Based on a fixed-length standardized test reading passage, the following reading times (in minutes) and comprehension scores (based on a 100-point scale) were recorded.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	n	$\bar{y}$	s
Reading Time	5	7	15	12	8	7	10	11	9	13	10	6	11	8	10	8	7	6	11	8	20	9.10	2.573
Comprehension	60	76	76	90	81	75	95	98	88	73	90	66	91	83	100	85	76	69	91	78	20	82.05	10.88

a. What is the population about which inferences are being made?
b. Place a 95% confidence interval on the mean reading time for all incoming freshmen in the district.
c. Plot the reading time using a normal probability plot or boxplot. Do the data appear to be a random sample from a population having a normal distribution?
d. Provide an interpretation of the interval estimate in part (b).

5.37 Refer to Exercise 5.36. Using the reading comprehension data, is there significant evidence that the reading program would produce for incoming freshmen a mean comprehension score greater than 80, the statewide average for comparable students during the previous year? Determine the level of significance for your test. Interpret your findings.

5.38 Refer to Exercise 5.36.

a. Does there appear to be a relationship between reading time and reading comprehension of the individual students? Provide a plot of the data to support your conclusion.
b. What are some weak points in this study relative to evaluating the potential of the reading improvement program? How would you redesign the study to overcome these weak points?

Bus. **5.39** A consumer testing agency wants to evaluate the claim made by a manufacturer of discount tires. The manufacturer claims that its tires can be driven at least 35,000 miles before wearing out. To determine the average number of miles that can be obtained from the manufacturer's tires, the agency randomly selects 60 tires from the manufacturer's warehouse and places the tires on 15 cars driven by test drivers on a 2-mile oval track. The number of miles driven (in thousands of miles) until the tires are determined to be worn out is given in the following table.

Car	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	n	$\bar{y}$	s
Miles Driven	25	27	35	42	28	37	40	31	29	33	30	26	31	28	30	15	31.47	5.04

a. Place a 99% confidence interval on the average number of miles driven, μ, prior to the tires wearing out.
b. Is there significant evidence ($\alpha = .01$) that the manufacturer's claim is false? What is the level of significance of your test? Interpret your findings.

5.40 Refer to Exercise 5.39.
a. Does the normality of the data appear to be valid?
b. How close to the true value were your bounds on the p-value?
c. Is there a contradiction between the interval estimate of μ and the conclusion reached by your test of the hypotheses?

Env. **5.41** The amount of sewage and industrial pollutants dumped into a body of water affects the health of the water by reducing the amount of dissolved oxygen available for aquatic life. Over a 2-month period, eight samples were taken from a river at a location 1 mile downstream from a sewage treatment plant. The amount of dissolved oxygen in the samples was determined and is reported in the following table. The current research asserts that the mean dissolved oxygen level must be at least 5.0 parts per million (ppm) for fish to survive.

Sample	1	2	3	4	5	6	7	8	n	$\bar{y}$	s
Oxygen (ppm)	5.1	4.9	5.6	4.2	4.8	4.5	5.3	5.2	8	4.95	.45

a. Place a 95% confidence on the mean dissolved oxygen level during the 2-month period.
b. Using the confidence interval from part (a), does the mean oxygen level appear to be less than 5 ppm?
c. Test the research hypothesis that the mean oxygen level is less than 5 ppm. What is the level of significance of your test? Interpret your findings.

Env. **5.42** A dealer in recycled paper places empty trailers at various sites. The trailers are gradually filled by individuals who bring in old newspapers and magazines and are picked up on several schedules. One such schedule involves pickup every second week. This schedule is desirable if the average amount of recycled paper is more than 1,600 cubic feet per 2-week period. The dealer's records for 18 2-week periods show the following volumes (in cubic feet) at a particular site:

1,660	1,820	1,590	1,440	1,730	1,680	1,750	1,720	1,900
1,570	1,700	1,900	1,800	1,770	2,010	1,580	1,620	1,690

$\bar{y} = 1{,}718.3$ and $s = 137.8$

a. Assuming the 18 2-week periods are fairly typical of the volumes throughout the year, is there significant evidence that the average volume μ is greater than 1,600 cubic feet?
b. Place a 95% confidence interval on μ.
c. Compute the p-value for the test statistic. Is there strong evidence that μ is greater than 1,600?

5.8 Inferences About μ When the Population Is Nonnormal and n Is Small: Bootstrap Methods

5.43 Refer to Exercise 5.36.
- **a.** Use a computer program to obtain 10,000 bootstrap samples from the 20 comprehension scores. Use these 10,000 samples to obtain the bootstrap p-value for the t test of H_a: $\mu > 80$.
- **b.** Compare the p-value from part (a) to the p-value obtained in Exercise 5.37.

5.44 Refer to Exercise 5.39.
- **a.** Use a computer program to obtain 10,000 bootstrap samples from the 15 sets of tire wear data. Use these 10,000 samples to obtain the bootstrap p-value for the t test of H_a: $\mu < 35$.
- **b.** Compare the p-value from part (a) to the p-value obtained in Exercise 5.39.

5.45 Refer to Exercise 5.41.
- **a.** Use a computer program to obtain 10,000 bootstrap samples from the eight oxygen levels. Use these 10,000 samples to obtain the bootstrap p-value for the t test of H_a: $\mu < 5$.
- **b.** Compare the p-value from part (a) to the p-value obtained in Exercise 5.41.

5.46 Refer to Exercise 5.42.
- **a.** Use a computer program to obtain 10,000 bootstrap samples from the 18 recycling volumes. Use these 10,000 samples to obtain the bootstrap p-value for the t test of H_a: $\mu > 1{,}600$.
- **b.** Compare the p-value from part (a) to the p-value obtained in Exercise 5.42.

5.9 Inferences About the Median

Basic **5.47** A random sample of 12 measurements is obtained from a population. Let M be the median for the population. The research study requires an estimate of M. The sample median is determined to be 37.8. The researchers want to assess a range of values for this point estimator.
- **a.** Display a 95% confidence interval on M by obtaining the values of $L_{\alpha/2}$ and $U_{\alpha/2}$.
- **b.** Obtain a 95% confidence interval on M using the large-sample approximations of $L_{\alpha/2}$ and $U_{\alpha/2}$. Compare the two confidence intervals.
- **c.** Provide reasons for the difference in the two confidence intervals.

Basic **5.48** A random sample of 50 measurements is obtained from a population. Let M be the median for the population. The research study requires an estimate of M. The sample median is determined to be 37.8. The researchers want to assess a range of values for this point estimator.
- **a.** Display a 95% confidence interval on M by obtaining the values of $L_{\alpha/2}$ and $U_{\alpha/2}$.
- **b.** Obtain a 95% confidence interval on M using the large-sample approximations of $L_{\alpha/2}$ and $U_{\alpha/2}$. Compare the two confidence intervals.
- **c.** Provide reasons for the difference in the two confidence intervals.

Basic **5.49** A researcher selects a random sample of 25 units from a population. Let M be the population median. Display the rejection region for an $\alpha = .01$ test that the population median is greater than 40.

Basic **5.50** Refer to Exercise 5.49.
- **a.** Display the rejection region for an $\alpha = .01$ test that the population median is greater than 40 using the large-sample approximation.
- **b.** Compare the rejection region from Exercise 5.49 to the rejection region in part (a). Provide reasons for the differences in the two regions.

Bus. **5.51** The amount of money spent on health care is an important issue for workers because many companies provide health insurance that only partially covers many medical procedures. The director of employee benefits at a midsize company wants to determine the amount spent on health

care by the typical hourly worker in the company. A random sample of 25 workers is selected, and the amounts they spent on their families' health care needs during the past year are given here.

400	345	248	1,290	398	218	197	342	208	223	531	172	4,321
143	254	201	3,142	219	276	326	207	225	123	211	108	

a. Graph the data using a boxplot or normal probability plot, and determine whether the population has a normal distribution.
b. Based on your answer to part (a), is the mean or the median cost per household a more appropriate measure of what the typical worker spends on health care needs?
c. Place a 95% confidence interval on the amount spent on health care by the typical worker. Explain what the confidence interval is telling us about the amount spent on health care needs.
d. Does the typical worker spend more than $400 per year on health care needs? Use $\alpha = .05$.

Gov. **5.52** Many states have attempted to reduce the blood-alcohol level at which a driver is declared to be legally drunk. There has been resistance to this change in the law by certain business groups who have argued that the current limit is adequate. A study was conducted to demonstrate the effect on reaction time of a blood-alcohol level of .1%, the current limit in many states. A random sample of 25 persons of legal driving age had their reaction times recorded in a standard laboratory test procedure before and after drinking a sufficient amount of alcohol to raise their blood alcohol to a .1% level. The difference (After − Before) in their reaction times in seconds was recorded as follows:

.01	.02	.04	.05	.07	.09	.11	.26	.27	.27	.28	.28	.29
.29	.30	.31	.31	.32	.33	.35	.36	.38	.39	.39	.40	

a. Graph the data and assess whether the population has a normal distribution.
b. Place a 99% confidence interval on both the mean and the median differences in reaction times of drivers who have a blood-alcohol level of .1%.
c. Is there sufficient evidence that a blood-alcohol level of .1% causes any increase in the mean reaction time?
d. Is there sufficient evidence that a blood-alcohol level of .1% causes any increase in the median reaction time?
e. Which summary of reaction time differences seems more appropriate, the mean or median? Justify your answer.

5.53 Refer to Exercise 5.52. The lobbyist for the business group has his expert examine the experimental equipment and determines that measurement errors may have been made when recording the reaction times. Unless the difference in reaction time is at least .25 seconds, the expert claims that the two times are essentially equivalent.

a. Is there sufficient evidence that the median difference in reaction times is greater than .25 seconds?
b. What other factors about the drivers are important in attempting to decide whether moderate consumption of alcohol affects reaction time?

Soc. **5.54** In an attempt to increase the amount of money people would receive at retirement from Social Security, the U.S. Congress during its 1999 session debated whether a portion of Social Security funds should be invested in the stock market. Advocates of mutual stock funds reassured the public by stating that most mutual funds would provide a larger retirement income than the income currently provided by Social Security. The annual rates of return of two highly recommended mutual funds for the years 1989 through 1998 are given here. (The annual rate of return is defined as $(P_1 - P_0)/P_0$, where P_0 and P_1 are the prices of the fund at the beginning and end of the year, respectively.)

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Fund A	25.4	17.1	−8.9	26.7	3.6	−8.5	−1.3	32.9	22.9	26.6
Fund B	31.9	−8.4	41.8	6.2	17.4	−2.1	30.5	15.8	26.8	5.7

a. For both fund A and fund B, estimate the mean and median annual rates of return, and construct a 95% confidence interval for each.

b. Which of the parameters, the mean or median, do you think best represents the annual rate of return for fund A and for fund B during the years 1989 through 1998? Justify your answer.

5.55 Refer to Exercise 5.54.

a. Is there sufficient evidence that the median annual rate of return for the two mutual funds is greater than 10%?

b. Is there sufficient evidence that the mean annual rate of return for the two mutual funds is greater than 10%?

5.56 Using the information in Table 5.8, answer the following questions.

a. If the population has a normal distribution, then the population mean and median are identical. Thus, either the mean or the median could be used to represent the center of the population. In this situation, why is the t test more appropriate than the sign test for testing hypotheses about the center of the distribution?

b. Suppose the population has a distribution that is highly skewed to the right. The researcher uses an $\alpha = .05$ t test to test hypotheses about the population mean. If the sample size is $n = 10$, will the probability of a Type I error for the test be .05? Justify your answer.

c. When testing hypotheses about the mean or median of a highly skewed population, the difference in power between the sign and t tests decreases as the size of $(M_a - M_0)$ increases. Verify this statement using the values in Table 5.8. Why do think this occurs?

d. When testing hypotheses about the mean or median of a lightly skewed population, the difference in power between the sign and t tests is much less than that for a highly skewed population distribution. Verify this statement using the values in Table 5.8. Why do you think this occurs?

Supplementary Exercises

Bus. **5.57** A Internet provider has implemented a new process for handling customer complaints. Based on a review of customer complaint data for the past 2 years, the mean time for handling a customer complain was 27 minutes. Three months after implementing the plan, a random sample of the records of 50 customers who had complaints produced the following response times. Use the 50 data values to determine if the new process has reduced the mean time to handle customer complaints.

32.3	26.9	25.4	32.9	27.7	32.2	24.8	20.5	30.4	21.3	25.9	27.1	19.2	28.4	18.0
33.1	31.1	21.9	33.4	24.3	25.5	29.6	32.7	21.3	31.8	27.6	17.4	26.9	18.9	28.6
23.5	21.6	20.1	30.9	26.8	28.7	24.6	21.5	21.9	28.3	24.1	28.9	29.8	27.1	23.8
25.3	30.7	27.2	19.0	30.0										

a. Estimate the mean time for handling a customer complaint under the new process using a 95% confidence interval.

b. Is there substantial evidence ($\alpha = .05$) that the new process has reduced the mean time to handle a customer complaint?

c. What is the population about which inferences from these data can be made?

Env. **5.58** The concentration of mercury in a lake has been monitored for a number of years. Measurements taken on a weekly basis yielded an average of 1.20 mg/m^3 (milligrams per cubic meter) with a standard deviation of .32 mg/m^3. Following an accident at a smelter on the shore of the lake, 15 measurements produced the following mercury concentrations.

1.60	1.77	1.61	1.08	1.07	1.79	1.34	1.07
1.45	1.59	1.43	2.07	1.16	0.85	2.11	

a. Give a point estimate of the mean mercury concentration after the accident.
b. Construct a 95% confidence interval on the mean mercury concentration after the accident. Interpret this interval.
c. Is there sufficient evidence that the mean mercury concentration has increased since the accident? Use $\alpha = .05$.
d. Assuming that the standard deviation of the mercury concentration is .32 mg/m^3, calculate the power of the test to detect mercury concentrations of 1.28, 1.32, 1.36, and 1.40.

Med. **5.59** In a standard dissolution test for tablets of a particular drug product, the manufacturer must obtain the dissolution rate for a batch of tablets prior to release of the batch. Suppose that the dissolution test consists of assays for 24 randomly selected individual 25 mg tablets. For each test, the tablet is suspended in an acid bath and then assayed after 30 minutes. The results of the 24 assays are given here.

19.5	19.7	19.7	20.4	19.2	19.5	19.6	20.8
19.9	19.2	20.1	19.8	20.4	19.8	19.6	19.5
19.3	19.7	19.5	20.6	20.4	19.9	20.0	19.8

a. Using a graphical display, determine whether the data appear to be a random sample from a normal distribution.
b. Estimate the mean dissolution rate for the batch of tablets, for both a point estimate and a 99% confidence interval.
c. Is there significant evidence that the batch of pills has a mean dissolution rate less than 20 mg (80% of the labeled amount in the tablets)? Use $\alpha = .01$.
d. Calculate the probability of a Type II error if the true dissolution rate is 19.6 mg.

Bus. **5.60** When an audit must be conducted that involves a tedious examination of a large inventory, the audit may be very costly and time consuming if each item in the inventory must be examined. In such situations, the auditor frequently obtains a random sample of items from the complete inventory and uses the results of an audit of the sampled items to check the validity of the company's financial statement. A large company's financial statement claims an inventory that averages $600 per item. The following data are the auditor's assessment of a random sample of 75 items from the company's inventory. The values resulting from the audit are rounded to the nearest dollar.

303	547	1,368	493	984	507	148	2,546	738	83	2	135	274	74	1,472
399	1,784	71	751	136	571	147	282	2,039	1,909	748	188	548	1	280
102	618	129	1,324	1,428	469	102	454	1,059	939	303	600	234	514	17
551	293	1,395	7	28	2	973	506	511	812	1,290	685	447	11	35
252	1,526	464	5	67	99	67	259	7	67	248	3,215	3	33	41

a. Estimate the mean value of an item in the inventory using a 95% confidence interval.
b. Is there substantial evidence ($\alpha = .01$) that the mean value of an item in the inventory is less than $600?
c. What is the target population for the above inferences?
d. Would normal distribution–based procedures be appropriate for answering the above questions?

Bus. **5.61** Over the past 5 years, the mean time for a warehouse to fill a buyer's order has been 25 minutes. Officials of the company believe that the length of time has increased recently, either due to a change in the workforce or due to a change in customer purchasing policies. The processing times (in minutes) were recorded for a random sample of 15 orders processed over the past month.

28	25	27	31	10
26	30	15	55	12
24	32	28	42	38

Do the data present sufficient evidence to indicate that the mean time to fill an order has increased?

Engin. **5.62** If a new process for mining copper is to be put into full-time operation, it must produce an average of more than 50 tons of ore per day. A 15-day trial period gave the results shown in the accompanying table.

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Yield (tons)	57.8	58.3	50.3	38.5	47.9	157.0	38.6	140.2	39.3	138.7	49.2	139.7	48.3	59.2	49.7

a. Estimate the typical amount of ore produced by the mine using both a point estimate and a 95% confidence interval.
b. Is there significant evidence that on a typical day the mine produces more than 50 tons of ore? Test by using $\alpha = .05$.

Env. **5.63** The board of health of a particular state was called to investigate claims that raw pollutants were being released into the river flowing past a small residential community. By applying financial pressure, the state was able to get the violating company to make major concessions toward the installation of a new water purification system. In the interim, different production systems were to be initiated to help reduce the pollution level of water entering the stream. To monitor the effect of the interim system, a random sample of 50 water specimens was taken throughout the month at a location downstream from the plant. If $\bar{y} = 5.0$ and $s = .70$, use the sample data to determine whether the mean dissolved oxygen count of the water (in ppm) is less than 5.2, the average reading at this location over the past year.

a. List the five parts of the statistical test, using $\alpha = .05$.
b. Conduct the statistical test and state your conclusion.

Env. **5.64** The search for alternatives to oil as a major source of fuel and energy will inevitably bring about many environmental challenges. These challenges will require solutions to problems in such areas as strip mining and many others. Let us focus on one. If coal is considered as a major source of fuel and energy, we will have to consider ways to keep large amounts of sulfur dioxide (SO_2) and particulates from getting into the air. This is especially important at large government and industrial operations. Here are some possibilities.

1. Build the smokestack extremely high.
2. Remove the SO_2 and particulates from the coal prior to combustion.
3. Remove the SO_2 from the gases after the coal is burned but before the gases are released into the atmosphere. This is accomplished by using a scrubber.

A new type of scrubber has been recently constructed and is set for testing at a power plant. Over a 15-day period, samples are obtained three times daily from gases emitted from the stack. The amounts of SO_2 emissions (in pounds per million BTU) are given here:

	Day														
Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6 A.M.	.158	.129	.176	.082	.099	.151	.084	.155	.163	.077	.116	.132	.087	.134	.179
2 P.M.	.066	.135	.096	.174	.179	.149	.164	.122	.063	.111	.059	.118	.134	.066	.104
10 P.M.	.128	.172	.106	.165	.163	.200	.228	.129	.101	.068	.100	.119	.125	.182	.138

a. Estimate the average amount of SO_2 emissions during each of the three time periods using 95% confidence intervals.
b. Does there appear to be a significant difference in the average amounts of SO_2 emissions over the three time periods?
c. Combining the data over the entire day, is the average amount of SO_2 emissions using the new scrubber less than .145, the average daily value for the old scrubber?

Soc. **5.65** As part of an overall evaluation of training methods, an experiment was conducted to determine the average exercise capacity of healthy male army inductees. To do this, each male in a random sample of 35 healthy army inductees exercised on a bicycle ergometer (a device for measuring work done by the muscles) under a fixed workload until he tired.

Blood pressure, pulse rate, and other indicators were carefully monitored to ensure that no one's health was in danger. The exercise capacities (mean time, in minutes) for the 35 inductees are listed here.

23	19	36	12	41	43	19
28	14	44	15	46	36	25
35	25	29	17	51	33	47
42	45	23	29	18	14	48
21	49	27	39	44	18	13

a. Use these data to construct a 95% confidence interval for μ, the average exercise capacity for healthy male inductees. Interpret your findings.
b. How would your interval change using a 99% confidence interval?

5.66 Using the data in Exercise 5.65, determine the number of sample observations that would be required to estimate μ to within 1 minute, using a 95% confidence interval.

H.R. **5.67** Faculty members in a state university system who resign within 10 years of initial employment are entitled to receive the money paid into a retirement system, plus 4% per year. Unfortunately, experience has shown that the state is extremely slow in returning this money. Concerned about such a practice, a local teachers' organization decides to investigate. For a random sample of 50 employees who resigned from the state university system over the past 5 years, the average time between the termination date and reimbursement was 75 days, with a standard deviation of 15 days. Use the data to estimate the mean time to reimbursement, using a 95% confidence interval.

5.68 Refer to Exercise 5.67. After a confrontation with the teachers' union, the state promised to make reimbursements within 60 days. Monitoring of the next 40 resignations yields an average of 58 days, with a standard deviation of 10 days. If we assume that these 40 resignations represent a random sample of the state's future performance, estimate the mean reimbursement time using a 99% confidence interval.

Bus. **5.69** Improperly filled orders are a costly problem for mail-order houses. To estimate the mean loss per incorrectly filled order, a large firm plans to sample n incorrectly filled orders and to determine the added cost associated with each one. The firm estimates that the added cost is between \$40 and \$400. How many incorrectly filled orders must be sampled to estimate the mean additional cost using a 95% confidence interval of width \$20?

Engin. **5.70** The recipe for producing a high-quality cement specifies that the required percentage of SiO_2 is 6.2%. A quality control engineer evaluates this specification weekly by randomly selecting samples from $n = 20$ batches on a daily basis. On a given day, she obtained the following values:

1.70	9.86	5.44	4.28	4.59	8.76	9.16	6.28	3.83	3.17
5.98	2.77	3.59	3.17	8.46	7.76	5.55	5.95	9.56	3.58

a. Estimate the mean percentage of SiO_2 using a 95% confidence interval.
b. Evaluate whether the percentage of SiO_2 is different from the value specified in the recipe using an $\alpha = .05$ test of hypotheses.
c. Produce a plot to determine if the procedures you used in parts (a) and (b) were valid.

5.71 Refer to Exercise 5.70.
a. Estimate the median percentage of SiO_2 using a 95% confidence interval.
b. Evaluate whether the median percentage of SiO_2 is different from 6.2% using an $\alpha = .05$ test of hypotheses.

5.72 Refer to Exercise 5.70. Generate 9,999 bootstrap samples from the 20 SiO_2 percentages.
a. Construct a 95% bootstrap confidence interval on the mean SiO_2 percentage. Compare this interval to the interval obtained in Exercise 5.70(a).
b. Obtain the bootstrap p-value for testing whether the mean percentage of SiO_2 differs from 6.2%. Compare this value to the p-value for the test in Exercise 5.70(b).
c. Why is there such a good agreement between the t-based and bootstrap values in parts (a) and (b)?

Med. **5.73** A medical team wants to evaluate the effectiveness of a new drug that has been proposed for people with high intraocular pressure (IOP). Prior to running a full-scale clinical trial of the drug, a pilot test was run using 10 patients with high IOP values. The $n = 10$ patients had a mean decrease in IOP of $\bar{y} = 15.2$ mm Hg with a standard deviation of the 10 IOPs equal to $s = 9.8$ mm Hg after 15 weeks of using the drug. Determine the appropriate sample size for an $\alpha = .01$ test to have at most a .10 probability of failing to detect at least a 4 mm Hg decrease in the mean IOP.

Gov. **5.74** A federal regulatory agency is investigating an advertised claim that a certain device can increase the gasoline mileage of cars (mpg). Ten such devices are purchased and installed in cars belonging to the agency. Gasoline mileage for each of the cars is recorded both before and after installation. The data are recorded here.

	Car												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	***n***	$\bar{x}$	***s***
Before (mpg)	19.1	29.9	17.6	20.2	23.5	26.8	21.7	25.7	19.5	28.2	10	23.22	4.25
After (mpg)	25.8	23.7	28.7	25.4	32.8	19.2	29.6	22.3	25.7	20.1	10	25.33	4.25
Change (mpg)	6.7	−6.2	11.1	5.2	9.3	−7.6	7.9	−3.4	6.2	−8.1	10	2.11	7.54

Place 90% confidence intervals on the average mpg for both the before and the after phases of the study. Interpret these intervals. Does it appear that the device will significantly increase the average mileage of cars?

5.75 Refer to Exercise 5.74.

a. The cars in the study appear to have grossly different mileages before the devices were installed. Use the change data to test whether there has been a significant gain in mileage after the devices were installed. Use $\alpha = .05$.

b. Construct a 90% confidence interval for the mean change in mileage. On the basis of this interval, can one reject the hypothesis that the mean change is either zero or negative? (Note that the two-sided 90% confidence interval corresponds to a one-tailed $\alpha = .05$ test by using this decision rule: Reject H_0: $\mu \geq \mu_0$ if μ_0 is greater than the upper limit of the confidence interval.)

5.76 Refer to Exercise 5.74.

a. Calculate the probability of a Type II error for several values of μ_c, the average change in mileage. How do these values affect the conclusion you reached in Exercise 5.75?

b. Suggest some changes in the way in which this study in Exercise 5.74 was conducted.

CHAPTER 6

Inferences Comparing Two Population Central Values

6.1 Introduction and Abstract of Research Study

The inferences we have made so far have concerned a parameter from a single population. Quite often we are faced with an inference involving a comparison of parameters from different populations. We might wish to compare the mean corn crop yields for two different varieties of corn, the mean annual incomes for two ethnic groups, the mean nitrogen contents of two different lakes, or the mean lengths of time between administration and eventual relief for two different antivertigo drugs.

In many sampling situations, we will select independent random samples from two populations to compare the populations' parameters. The statistics used to make these inferences will, in many cases, be the differences between the corresponding sample statistics. Suppose we select independent random samples of n_1 observations from one population and n_2 observations from a second population. We will use the difference between the sample means, $(\bar{y}_1 - \bar{y}_2)$, to make an inference about the difference between the population means, $(\mu_1 - \mu_2)$.

The following theorem will help in finding the sampling distribution for the difference between sample statistics computed from independent random samples.

THEOREM 6.1

If two independent random variables y_1 and y_2 are normally distributed with means and variances (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively, the difference between the random variables is normally distributed with mean $(\mu_1 - \mu_2)$ and variance $(\sigma_1^2 + \sigma_2^2)$. Similarly, the sum $(y_1 + y_2)$ of the random variables is normally distributed with mean $(\mu_1 + \mu_2)$ and variance $(\sigma_1^2 + \sigma_2^2)$.

Theorem 6.1 can be applied directly to find the sampling distribution of the difference between two independent sample means or two independent sample proportions. The Central Limit Theorem (discussed in Chapter 4) implies that if two random samples of sizes, n_1 and n_2, are independently selected from two populations, 1 and 2, then where n_1 and n_2 are large, the sampling distributions of $\bar{y}_1$ and $\bar{y}_2$ will be approximately normal with means and variances $(\mu_1, \sigma_1^2/n_1)$ and $(\mu_2, \sigma_2^2/n_2)$, respectively. Consequently, because $\bar{y}_1$ and $\bar{y}_2$ are independent, normally distributed random variables, it follows from Theorem 6.1 that the sampling distribution for the difference in the sample means, $(\bar{y}_1 - \bar{y}_2)$, is approximately normal with a mean of

$$\mu_{\bar{y}_1 - \bar{y}_2} = \mu_1 - \mu_2$$

a variance of

$$\sigma^2_{\bar{y}_1 - \bar{y}_2} = \sigma^2_{\bar{y}_1} + \sigma^2_{\bar{y}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

and a standard error of

$$\sigma_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Properties of the Sampling Distribution for the Difference Between Two Sample Means, $(\bar{y}_1 - \bar{y}_2)$

1. The sampling distribution of $(\bar{y}_1 - \bar{y}_2)$ is approximately normal for large samples.
2. The mean of the sampling distribution, $\mu_{\bar{y}_1 - \bar{y}_2}$, is equal to the difference between the population means, $(\mu_1 - \mu_2)$.
3. The standard error of the sampling distribution is

$$\sigma_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The sampling distribution of the difference between two independent, normally distributed sample means is shown in Figure 6.1.

The sampling distribution for the difference between two sample means, $(\bar{y}_1 - \bar{y}_2)$, can be used to answer the same types of questions as we asked about the sampling distribution for $\bar{y}$ in Chapter 4. Because sample statistics are used to make inferences about corresponding population parameters, we can use the sampling distribution of a statistic to calculate the probability that the statistic will

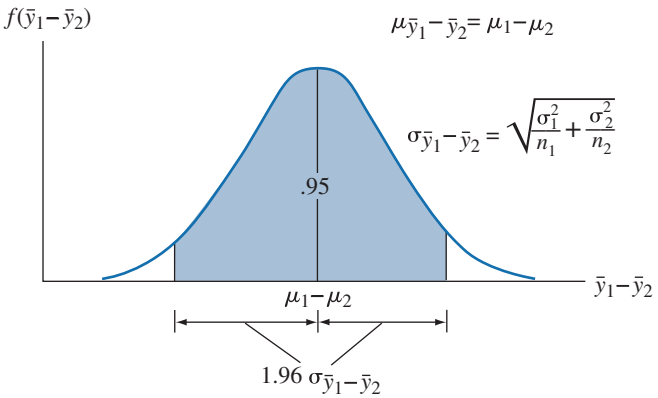


FIGURE 6.1
Sampling distribution for the difference between two sample means

be within a specified distance of the population parameter. For example, we could use the sampling distribution of the difference in sample means to calculate the probability that $(\bar{y}_1 - \bar{y}_2)$ will be within a specified distance of the unknown difference in population means, $(\mu_1 - \mu_2)$. Inferences (estimations or tests) about $(\mu_1 - \mu_2)$ will be discussed in succeeding sections of this chapter.

Abstract of Research Study: Effects of an Oil Spill on Plant Growth

On January 7, 1992, an underground oil pipeline ruptured and caused the contamination of a marsh along the Chiltipin Creek in San Patricio County, Texas. The cleanup process consisted of a number of procedures, including vacuuming the spilled oil, burning the contaminated region in the marsh to remove the remaining oil, and then planting native plants in the contaminated region. Federal regulations require the company responsible for the oil spill to document that the contaminated region has been restored to its prespill condition. To evaluate the effectiveness of the cleanup process and, in particular, to study the residual effects of the oil spill on the flora, researchers designed a study of plant growth 1 year after the burning. In an unpublished Texas A&M University dissertation, ***Newman (1998)*** describes the researchers' plan for evaluating the effect of the oil spill on *Distichlis spicata,* a flora of particular importance to the area of the spill.

After holding lengthy discussions, reading the relevant literature, and searching many data bases about similar sites and flora, the researchers found there was no specific information on the flora in this region prior to the oil spill. They determined that the flora parameters of interest were the average *Distichlis spicata* density μ after burning the spill region, the variability σ in flora density, and the proportion π of the spill region in which the flora density was essentially zero. Since there was no relevant information on flora density in the spill region prior to the spill, it was necessary to evaluate the flora density in unaffected areas of the marsh to determine whether the plant density had changed after the oil spill. The researchers located several regions that had not been contaminated by the oil spill. The spill region and the unaffected regions were divided into tracts of nearly the same size. The number of tracts needed in the study was determined by specifying how accurately the parameters μ, σ, and π needed to be estimated in order to achieve a level of precision as specified by the width of 95% confidence intervals and by the power of tests of hypotheses. From these calculations and within budget and time limitations, it was decided that 40 tracts from both the spill and the unaffected areas would be used in the study.

Forty tracts of exactly the same size were randomly selected in these locations, and the *Distichlis spicata* density was recorded. Similar measurements were taken within the spill area of the marsh. The data are presented in Section 6.7.

From the data, summary statistics were computed in order to compare the two sites. The average flora density in the control sites is $\bar{y}_{\text{Con}} = 38.48$ with a standard deviation of $s_{\text{Con}} = 16.37$. The sites within the spill region have an average density of $\bar{y}_{\text{Spill}} = 26.93$ with a standard deviation of $s_{\text{Spill}} = 9.88$. Thus, the control sites have a larger average flora density and a greater variability in flora density than do the sites within the spill region. Whether these observed differences in flora density reflect similar differences in all the sites and not just the ones included in the study will require a statistical analysis of the data. We will discuss the construction of confidence intervals and statistical tests about the differences between μ_{Con} and μ_{Spill} in Section 6.7. The estimation and testing of the population standard deviations, σs, and population proportions, πs, will be the topic of Chapters 7 and 10. At the end of this chapter, we will provide an analysis of the data sets to determine if there is evidence that the conditions in the spill area have been returned to a state that is similar to its prespill condition.

6.2 Inferences About $\mu_1 - \mu_2$: Independent Samples

In situations where we are making inferences about $\mu_1 - \mu_2$ based on random samples independently selected from two populations, we will consider three cases:

Case 1. Both population distributions are normally distributed with $\sigma_1 = \sigma_2$.
Case 2. Both sample sizes, n_1 and n_2, are large.
Case 3. The sample sizes, n_1 or n_2, are small, and the population distributions are nonnormal.

In this section, we will consider the situation in which we are independently selecting random samples from two populations that have normal distributions with different means, μ_1 and μ_2. The data will be summarized into the statistics: sample means $\bar{y}_1$ and $\bar{y}_2$ and sample standard deviations s_1 and s_2. We will compare the two populations by constructing appropriate graphs, confidence intervals for $\mu_1 - \mu_2$, and tests of hypotheses concerning the difference $\mu_1 - \mu_2$.

A logical point estimate for the difference in population means is the sample difference $\bar{y}_1 - \bar{y}_2$. The standard error for the difference in sample means is more complicated than for a single sample mean, but the confidence interval has the same form: point estimate $\pm t_{\alpha/2}$ (standard error). A general confidence interval for $\mu_1 - \mu_2$ with a confidence level of $(1 - \alpha)$ is given here for the situations $\sigma_1 = \sigma_2$.

Confidence Interval for $\mu_1 - \mu_2$, Independent Samples Equal Variances

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad \text{and} \quad \text{df} = n_1 + n_2 - 2$$

The sampling distribution of $\bar{y}_1 - \bar{y}_2$ is a normal distribution with standard deviation

$$\sigma_{\bar{y}_1-\bar{y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

because we require that the two populations have the same standard deviation σ. If we knew the value of σ, then we would use $z_{\alpha/2}$ in the formula for the confidence interval. Because σ is unknown in most cases, we must estimate its value. This estimate is denoted by s_p and is determined by combining (pooling) the two independent estimates of σ, s_1, and s_2. In fact, $\boldsymbol{s_P^2}$ **is a weighted average** of the sample variances s_1^2 and s_2^2. We have to estimate the standard deviation of the point estimate of $\mu_1 - \mu_2$, so we must use the percentile from the t distribution, $t_{\alpha/2}$, in place of the normal percentile, $z_{\alpha/2}$. The degrees of freedom for the t-percentile are $\text{df} = n_1 + n_2 - 2$ because we have a total of $n_1 + n_2$ data values and two parameters, μ_1 and μ_2, that must be estimated prior to estimating the standard deviation σ. Remember that we use $\bar{y}_1$ and $\bar{y}_2$ in place of μ_1 and μ_2, respectively, in the formulas for s_1^2 and s_2^2.

s_p^2, a weighted average

Recall that we are assuming that the two populations from which we draw the samples have normal distributions with a common variance σ^2. If the confidence interval presented was valid only when these assumptions were met exactly, the estimation procedure would be of limited use. Fortunately, the confidence coefficient remains relatively stable if both distributions are mound-shaped and the sample sizes are approximately equal. For those situations in which these conditions do not hold, we will discuss alternative procedures in this section and in Section 6.3.

EXAMPLE 6.1

Company officials were concerned about the length of time a particular drug product retained its potency. A random sample of $n_1 = 10$ bottles of the product was drawn from the production line and analyzed for potency.

A second sample of $n_2 = 10$ bottles was obtained and stored in a regulated environment for a period of 1 year. The readings obtained from each sample are given in Table 6.1.

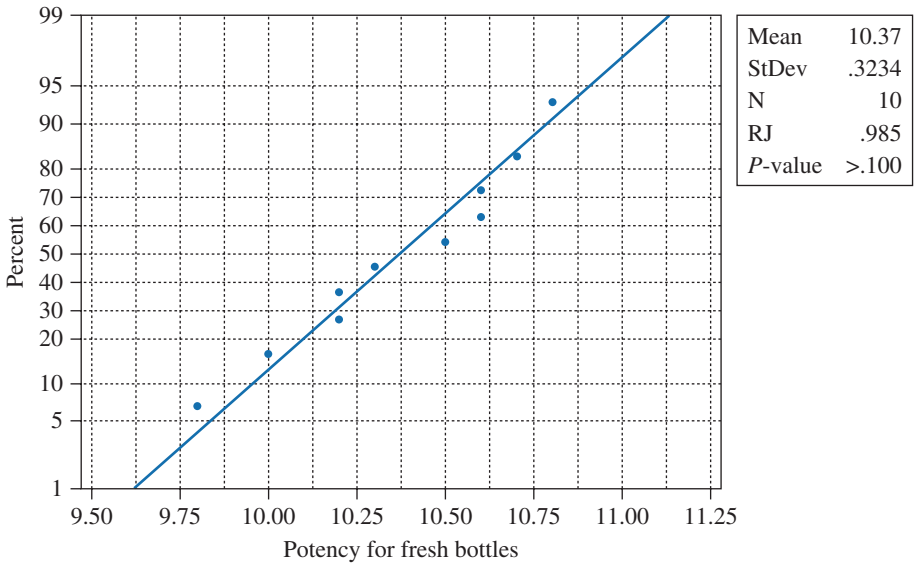
TABLE 6.1
Potency reading for two samples

Fresh		Stored	
10.2	10.6	9.8	9.7
10.5	10.7	9.6	9.5
10.3	10.2	10.1	9.6
10.8	10.0	10.2	9.8
9.8	10.6	10.1	9.9

Suppose we let μ_1 denote the mean potency for all bottles that might be sampled coming off the production line and let μ_2 denote the mean potency for all bottles that may be retained for a period of 1 year. Estimate $\mu_1 - \mu_2$ using a 95% confidence interval.

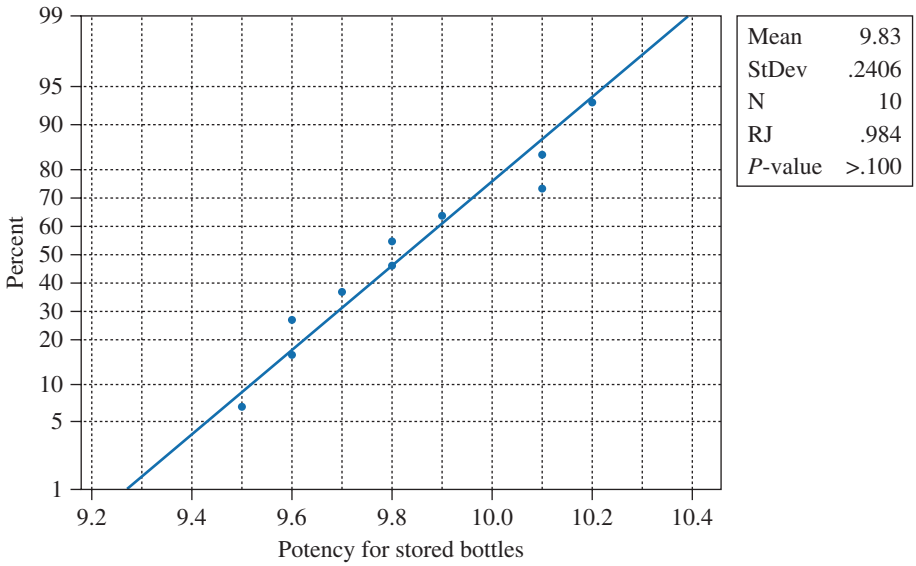
Solution The potency readings for the fresh and stored bottles are plotted in Figures 6.2(a) and (b) in normal probability plots to assess the normality assumption. We find that the plotted points in both plots fall very close to a straight line,

FIGURE 6.2(a)
Normal probability plot: potency of fresh bottles



(a)

FIGURE 6.2(b)
Normal probability plot: potency of stored bottles



(b)

and, hence, the normality condition appears to be satisfied for both types of bottles. The summary statistics for the two samples are presented next.

Fresh Bottles	**Stored Bottles**
$n_1 = 10$	$n_2 = 10$
$\bar{y}_1 = 10.37$	$\bar{y}_2 = 9.83$
$s_1 = 0.3234$	$s_2 = 0.2406$

In Chapter 7, we will provide a test of equality for two population variances. However, for the above data, the computed sample standard deviations are approximately equal considering the small sample sizes. Thus, the conditions required to construct a confidence interval on $\mu_1 - \mu_2$—that is, normality, equal

variances, and independent random samples—appear to be satisfied. The estimate of the common standard deviation σ is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{9(.3234)^2 + 9(.2406)^2}{18}} = .285$$

From Table 2 in the Appendix, the t-percentile based on df $= n_1 + n_2 - 2 = 18$ and $\alpha = .025$ is 2.101. A 95% confidence interval for the difference in mean potencies is

$$(10.37 - 9.83) \pm 2.101(.285)\sqrt{1/10 + 1/10}$$

$$.54 \pm .268 = (.272, .808)$$

We estimate that the difference in mean potencies for the bottles from the production line and those stored for 1 year, $\mu_1 - \mu_2$, lies in the interval .272 to .808. Company officials would then have to evaluate whether a decrease in mean potency of a size between .272 and .808 would have a practical impact on the useful potency of the drug. ■

EXAMPLE 6.2

During the past 20 years, the domestic automobile industry has been repeatedly challenged by consumer groups to raise the quality of their cars to the level of comparably priced imports. An automobile industry association decides to compare the mean repair costs of two models: a popular full-sized imported car and a widely purchased full-sized domestic car. The engineering firm hired to run the tests proposes driving the vehicles at a speed of 30 mph into a concrete barrier. The costs of the repairs to the vehicles will then be assessed. To account for variation in the damage to the vehicles, it is decided to use 10 imported cars and 10 domestic cars. After completing the crash testing, it was determined that the speed of one of the imported cars had exceeded 30 mph and thus was not a valid test run. Because of budget constraints, it was decided not to run another crash test using a new imported vehicle. The data, recorded in thousands of dollars, produced sample means and standard deviations as shown in Table 6.2. Use these data to construct a 95% confidence interval on the difference in mean repair costs, $(\mu_{\text{domestic}} - \mu_{\text{imported}}) = (\mu_1 - \mu_2)$.

TABLE 6.2
Summary of repair cost data for Example 6.2

	Domestic	Imported
Sample Size	10	9
Sample Mean	8.27	6.78
Sample Standard Deviation	2.956	2.565

Solution A normal probability of the data for each of the two samples suggests that the populations of damage repairs are nearly normally distributed. Also, considering the very small sample sizes, the closeness in size of the sample standard deviations would not indicate a difference in the population standard deviations; that is, it is appropriate to conclude that $\sigma_1 \approx \sigma_2 = \sigma$. Thus, the conditions necessary for applying the pooled t-based confidence intervals would appear to be satisfied.

The difference in sample means is

$$\bar{y}_1 - \bar{y}_2 = 8.27 - 6.78 = 1.49$$

The estimate of the common standard deviation in repair costs σ is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(10 - 1)(2.956)^2 + (9 - 1)(2.565)^2}{10 + 9 - 2}} = 2.778$$

The t-percentile for $\alpha/2 = .025$ and df $= 10 + 9 - 2 = 17$ is given in Table 2 of the Appendix as 2.110. A 95% confidence interval for the difference in mean repair costs is given here.

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Substituting the values from the repair cost study into the formula, we obtain

$$1.49 \pm 2.110(2.778)\sqrt{\frac{1}{10} + \frac{1}{9}} = 1.49 \pm 2.69 = (-1.20, 4.18)$$

Thus, we estimate the difference in mean repair costs between particular brands of domestic and imported cars tested to lie somewhere between -1.20 and 4.18. If we multiply these limits by \$1,000, the 95% confidence interval for the difference in mean repair costs is $-$\$1,200 to \$4,180. This interval includes both positive and negative values for $\mu_1 - \mu_2$, so we are unable to determine whether the mean repair cost for domestic cars is larger or smaller than the mean repair cost for imported cars. ■

We can also test a hypothesis about the difference between two population means. As with any test procedure, we begin by specifying a research hypothesis for the difference in population means. Thus, we might, for example, specify that the difference $\mu_1 - \mu_2$ is greater than some value D_0. (*Note:* D_0 will often be 0.) The entire test procedure is summarized here.

A Statistical Test for $\mu_1 - \mu_2$, Independent Samples, Equal Variances

The assumptions under which the test will be valid are the same as were required for constructing the confidence interval on $\mu_1 - \mu_2$: population distributions are normal with equal variances, and the two random samples are independent.

H_0:
1. $\mu_1 - \mu_2 \leq D_0$ (D_0 is a specified value, often 0)
2. $\mu_1 - \mu_2 \geq D_0$
3. $\mu_1 - \mu_2 = D_0$

H_a:
1. $\mu_1 - \mu_2 > D_0$
2. $\mu_1 - \mu_2 < D_0$
3. $\mu_1 - \mu_2 \neq D_0$

T.S.: $$t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

R.R.: For a level α, Type I error rate and with df $= n_1 + n_2 - 2$,
1. Reject H_0 if $t \geq t_\alpha$.
2. Reject H_0 if $t \leq -t_\alpha$.
3. Reject H_0 if $|t| \geq t_{\alpha/2}$.

Check assumptions and draw conclusions.

EXAMPLE 6.3

An experiment was conducted to evaluate the effectiveness of a treatment for tapeworm in the stomachs of sheep. A random sample of 24 worm-infected lambs of approximately the same age and health was randomly divided into two groups. Twelve of the lambs were injected with the drug, and the remaining 12 were left untreated. After a 6-month period, the lambs were slaughtered, and the worm counts recorded are listed in Table 6.3:

TABLE 6.3
Sample data for treated and untreated sheep

Drug-Treated Sheep	18	43	28	50	16	32	13	35	38	33	6	7
Untreated Sheep	40	54	26	63	21	37	39	23	48	58	28	39

a. Is there significant evidence that the untreated lambs have a mean tapeworm count that is more than five units greater than the mean count for the treated lambs? Use an $\alpha = .05$ test.
b. What is the level of significance for this test?
c. Place a 95% confidence interval on $\mu_1 - \mu_2$ to assess the size of the difference in the two means.

Solution

a. Boxplots of the worm counts for the treated and untreated lambs are displayed in Figure 6.3. From the plots, we can observe that the data for the untreated lambs are symmetric with no outliers and the data for the treated lambs are slightly skewed to the left with no outliers. Also, the widths of the two boxes are approximately equal. Thus, the condition that the population distributions are normal with equal variances appears to be satisfied. The condition of independence of the worm counts both between and within the two groups is evaluated by considering how the lambs were selected, assigned to the two groups, and cared for during the 6-month experiment. Because the 24 lambs were randomly selected from a representative herd of infected lambs, were randomly assigned to the treated and untreated groups, and were properly separated and cared for during the 6-month period of the experiment, the 24 worm counts are presumed to be independent random samples from the two populations. Finally, we can observe from the boxplots that the untreated lambs appear to have higher worm counts than the treated lambs because the median line is higher for the untreated group. The following test confirms our observation. The data for the treated and untreated sheep are summarized next.

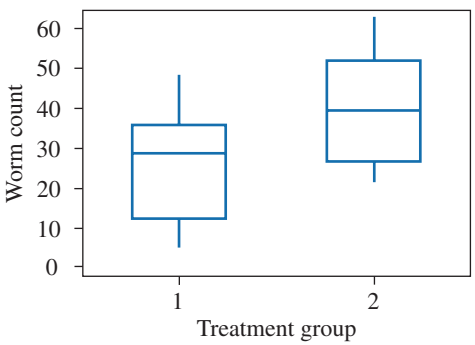


FIGURE 6.3
Boxplots of worm counts for treated (1) and untreated (2) sheep

Drug-Treated Lambs	Untreated Lambs
$n_2 = 12$	$n_1 = 12$
$\bar{y}_2 = 26.58$	$\bar{y}_1 = 39.67$
$s_2 = 14.36$	$s_1 = 13.86$

The sample standard deviations are of a similar size, so from this and from our observation from the boxplot, the pooled estimate of the common population standard deviation σ is now computed:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{11(13.86)^2 + 11(14.36)^2}{22}} = 14.11$$

The test procedure for evaluation of the research hypothesis that the untreated lambs have a mean tapeworm count (μ_1) that is more than five units greater than the mean count (μ_2) of the treated lambs is as follows:

H_0: $\mu_1 - \mu_2 \leq 5$ (drug does not reduce the mean tapeworm count by more than 5 units)

H_a: $\mu_1 - \mu_2 > 5$ (drug does reduce the mean tapeworm count by more than 5 units)

$$\text{T.S.: } t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(39.67 - 26.58) - 5}{14.11\sqrt{\frac{1}{12} + \frac{1}{12}}} = 1.404$$

R.R.: Reject H_0 if $t \geq 1.717$, where 1.717 is the value from Table 2 in the Appendix for a critical t-value with $\alpha = .05$ and df $= n_1 + n_2 - 2 = 22$.

Conclusion: Because the observed value of $t = 1.404$ is less than 1.717 and hence is not in the rejection region, there is insufficient evidence to conclude that the drug treatment reduces the mean tapeworm count by five or more units.

b. Using Table 2 in the Appendix with $t = 1.404$ and df $= 22$, we can bound the level of significance (p-value) in the range $.05 < p\text{-value} < .10$.

Using the R function ***pt*(t_c, df)**, which calculates $P(t \leq t_c)$, we can obtain the p-value for the calculated value of the T.S., $t_c = 1.404$.

p-value $= P(t \geq 1.404) = 1 - P(t \leq 1.404) = 1 - \textbf{\textit{pt}}(1.404, 22) = 1 - .913 = .087$

c. A 95% confidence interval on $\mu_1 - \mu_2$ provides the experimenter with an estimate of the size of the reduction in mean tapeworm count obtained by using the drug. This interval can be computed as follows:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{.025}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(39.67 - 26.58) \pm (2.074)(14.11)\sqrt{\frac{1}{12} + \frac{1}{12}} = 13.09 \pm 11.95 = (1.14, 25.4)$$

Thus, we are 95% certain that the reduction in mean tapeworm count through the use of the drug is between 1.1 and 25.0 worms. The confidence interval contains values that are less than 5, which is consistent with our conclusions. ■

The confidence interval and test procedures for comparing two population means presented in this section require three conditions to be satisfied. The first and most critical condition is that the two random samples are independent. Practically, we mean that the two samples are randomly selected from two distinct populations and that the elements of one sample are statistically independent of those of the second sample. Two types of dependencies (data are not independent) commonly occur in experiments and studies. The data may have a *cluster effect,* which often results when the data have been collected in subgroups. For example, 50 children are selected from five different classrooms for an experiment to compare the effectiveness of two tutoring techniques. The children are randomly assigned to one of the two techniques. Because children from the same classroom have a common teacher and hence may tend to be more similar in their academic achievement than children from different classrooms, the condition of independence between participants in the study may be lacking.

A second type of dependence is the result of serial or spatial correlation. When measurements are taken over time, observations that are closer together in time tend to be *serially correlated*—that is, more similar than observations collected at greatly different times. A similar dependence occurs when the data are collected at different locations—for example, water samples taken at various locations in a lake to assess whether a chemical plant is discharging pollutants into the lake. Measurements that are physically closer to each other are more likely to be similar than measurements taken farther apart. This type of dependence is *spatial correlation.* When the data are dependent, the procedures based on the t distribution produce confidence intervals having coverage probabilities different from the intended values and tests of hypotheses having Type I error rates different from the stated values. There are appropriate statistical procedures for handling this type of data, but they are more advanced. A book on longitudinal or repeated measures data analysis or the analysis of spatial data can provide the details for the analysis of dependent data.

When the population distributions are either very heavily tailed or highly skewed, the coverage probability for confidence intervals and the level and power of the t test will differ greatly from the stated values. A nonparametric alternative to the t test is presented in the next section; this test does not require normality.

The third assumption is that the two population variances, σ_1^2 and σ_2^2, are equal. In Chapter 7, a formal test of the equality of the two variances, named the F test, will be presented. However, the F test is not very reliable if the population distributions are not close to a normal distribution. Thus, use of the F test is not recommended in deciding whether the equal variance t-procedures are appropriate. If there is evidence in the data that the two variances are considerably different, then alternatives to the equal-variance t test should be implemented. In particular, if one of the variances is at least four times the other (e.g., $\sigma_1^2 = 4\sigma_2^2$), then the equal-variance t test and confidence intervals should not be used.

To illustrate the effect of unequal variances, a computer simulation was performed in which two independent random samples were generated from normal populations having the same means but unequal variances: $\sigma_1 = k\sigma_2$ with $k = .25$, .5, 1, 2, and 4. For each combination of sample sizes and standard deviations, 1,000 simulations were run. For each simulation, a level .05 test was conducted. The proportions of the 1,000 tests that incorrectly rejected H_0 are presented in Table 6.4. If the pooled t test is unaffected by the unequal variances, we would expect the proportions to be close to .05, the intended level, in all cases.

From the results in Table 6.4, we can observe that when the sample sizes are equal, the proportion of Type I errors remains close to .05 (ranging from .042 to .065). When the sample sizes are different, the proportion of Type I errors

TABLE 6.4
The effect of unequal variances on the Type I error rates of the pooled t test

		$\sigma_1 = k\sigma_2$				
n_1	n_2	$k = .25$	.50	1	2	4
10	10	.065	.042	.059	.045	.063
10	20	.016	.017	.049	.114	.165
10	40	.001	.004	.046	.150	.307
15	15	.053	.043	.056	.060	.060
15	30	.007	.023	.066	.129	.174
15	45	.004	.010	.069	.148	.250

deviates greatly from .05. The more serious case occurs when the smaller sample size is associated with the larger variance. In this case, the error rates are much larger than .05. For example, when $n_1 = 10$, $n_2 = 40$, and $\sigma_1 = 4\sigma_2$, the error rate is .307. However, when $n_1 = 10, n_2 = 10$, and $\sigma_1 = 4\sigma_2$, the error rate is .063, much closer to .05. This is remarkable and provides a convincing argument to use equal sample sizes.

In the situation in which the sample variances (s_1^2 and s_2^2) suggest that $\sigma_1^2 \neq \sigma_2^2$, there is an approximate t test using the test statistic

$$t' = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Welch (1947) showed that the percentage points of a t distribution with modified degrees of freedom, known as Welch-Satterthwaite approximation, can be used to set the rejection region for t'. This approximate t test is summarized here.

Approximate t Test for Independent Samples, Unequal Variance

H_0: **1.** $\mu_1 - \mu_2 \leq D_0$ **2.** $\mu_1 - \mu_2 \geq D_0$ **3.** $\mu_1 - \mu_2 = D_0$

H_a: **1.** $\mu_1 - \mu_2 > D_0$ **2.** $\mu_1 - \mu_2 < D_0$ **3.** $\mu_1 - \mu_2 \neq D_0$

T.S.: $$t' = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

R.R.: For a level α, Type I error rate,

1. Reject H_0 if $t' \geq t_\alpha$

2. Reject H_0 if $t' \leq -t_\alpha$

3. Reject H_0 if $|t'| \geq t_{\alpha/2}$

with

$$\text{df} = \frac{(n_1 - 1)(n_2 - 1)}{(1 - c)^2(n_1 - 1) + c^2(n_2 - 1)} \quad \text{and} \quad c = \frac{s_1^2/n_1}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note: If the computed value of df is not an integer, *round down* to the nearest integer.

The test based on the t' statistic is sometimes referred to as the *separate-variance t test* because we use the separate sample variances s_1^2 and s_2^2 rather than a pooled sample variance.

When there is a large difference between σ_1 and σ_2, we must also modify the confidence interval for $\mu_1 - \mu_2$. The following formula is developed from the separate-variance t test.

Approximate Confidence Interval for $\mu_1 - \mu_2$, Independent Samples with $\sigma_1 \neq \sigma_2$

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where the t percentile has

$$\text{df} = \frac{(n_1 - 1)(n_2 - 1)}{(1 - c)^2(n_1 - 1) + c^2(n_2 - 1)} \qquad \text{with } c = \frac{s_1^2/n_1}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

EXAMPLE 6.4

The weekend athlete often incurs an injury due to not having the most appropriate or latest equipment. For example, tennis elbow is an injury that is the result of the stress encountered by the elbow when striking a tennis ball. There have been enormous improvements in the design of tennis rackets in the last 20 years. To investigate whether the new oversized racket delivers less stress to the elbow than does a more conventionally sized racket, a group of 45 tennis players of intermediate skill volunteered to participate in the study. Because there was no current information on the oversized rackets, an unbalanced design was selected. Thirty-three players were randomly assigned to use the oversized racket, and the remaining 12 players used the conventionally sized racket. The force on the elbow just after the impact of a forehand strike of a tennis ball was measured five times for each of the 45 tennis players. The mean force was then taken of the five force readings; the summary of these 45 force readings is given in Table 6.5.

TABLE 6.5
Summary of force readings for Example 6.4

	Oversized	Conventional
Sample Size	33	12
Sample Mean	25.2	33.9
Sample Standard Deviation	8.6	17.4

Use the information in Table 6.5 to test the research hypothesis that a tennis player would encounter a smaller mean force at the elbow using an oversized racket than he or she would encounter using a conventionally sized racket.

Solution A normal probability of the force data for each type of racket suggests that the two populations of forces are nearly normally distributed. That the sample standard deviation in the forces for the conventionally sized racket is more than double that for the oversized racket would indicate a difference in the population

standard deviations. Thus, it would not be appropriate to conclude that $\sigma_1 \approx \sigma_2$. The separate-variance t test was applied to the data. The test procedure for evaluating the research hypothesis that the oversized racket has a smaller mean force is as follows:

H_0: $\mu_1 \geq \mu_2$ (that is, oversized racket does not have smaller mean force)
H_a: $\mu_1 < \mu_2$ (that is, oversized racket has smaller mean force)

Writing the hypotheses in terms of $\mu_1 - \mu_2$ yields

$$H_0: \mu_1 - \mu_2 \geq 0 \text{ versus } H_a: \mu_1 - \mu_2 < 0$$

$$\text{T.S.: } t' = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(25.2 - 33.9) - 0}{\sqrt{\frac{(8.6)^2}{33} + \frac{(17.4)^2}{12}}} = -1.66$$

To compute the rejection region and p-value, we need to compute the approximate df for t':

$$c = \frac{s_1^2/n_1}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \frac{(8.6)^2/33}{\frac{(8.6)^2}{33} + \frac{(17.4)^2}{12}} = .0816$$

$$\text{df} = \frac{(n_1 - 1)(n_2 - 1)}{(1 - c^2)(n_1 - 1) + c^2(n_2 - 1)}$$

$$= \frac{(33 - 1)(12 - 1)}{(1 - .0816)^2(33 - 1) + (.0816)^2(12 - 1)} = 13.01$$

We round 13.01 down to 13.

Table 2 in the Appendix has the t-percentile for $\alpha = .05$ equal to 1.771. We can now construct the rejection region.

R.R.: For $\alpha = .05$ and df = 13, reject H_0 if $t' < -1.771$.

Because $t' = -1.66$ is not less than -1.771, we fail to reject H_0 and conclude that there is not significant evidence that the mean force of oversized rackets is smaller than the mean force of conventionally sized rackets. We can bound the p-value using Table 2 in the Appendix with df = l3. With $t' = -1.66$, we conclude $.05 < p\text{-value} < .10$. Using a software package, the p-value is computed to be .060. ■

The standard practice in many studies is to always use the pooled t test. To illustrate that this type of practice may lead to improper conclusions, we will conduct the pooled t test on the above data. The estimate of the common standard deviation in mean force σ is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(33 - 1)(8.6)^2 + (12 - 1)(17.4)^2}{33 + 12 - 2}} = 11.5104$$

$$\text{T.S.: } t' = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(25.2 - 33.9) - 0}{11.5104\sqrt{\frac{1}{33} + \frac{1}{12}}} = -2.24$$

The t-percentile for $\alpha = .05$ and df $= 33 + 12 - 2 = 43$ is given in Table 2 of the Appendix as 1.684 (for df $= 40$). We can now construct the rejection region.

R.R.: For $\alpha = .05$ and df $= 43$, reject H_0 if $t < -1.684$.

Because $t = -2.24$ is less than -1.684, we would reject H_0 and conclude that there is significant evidence that the mean force of oversized rackets is smaller than the mean force of conventionally sized rackets. Using a software package, the p-value is computed to be .015. Thus, an application of the pooled t test when there is strong evidence of a difference in variances would lead to a wrong conclusion concerning the difference in the two means.

Although we failed to determine that the mean force delivered by the oversized racket was statistically significantly lower than the mean force delivered by the conventionally sized racket, the researchers may be interested in the range of values for the difference in the mean forces of the two types of rackets. We will now estimate the size of the difference in the two mean forces, $\mu_1 - \mu_2$, using a 95% confidence interval.

Using df $= 13$, as computed previously, the t-percentile from Table 2 in the Appendix is $t_{\alpha/2} = t_{.025} = 2.160$. Thus, the confidence interval is given by the following calculations:

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 25.2 - 33.9 \pm 2.16\sqrt{\frac{(8.6)^2}{33} + \frac{(17.4)^2}{12}}$$

$$= -8.7 \pm 11.32$$

Thus, we are 95% confident that the difference in the mean forces is between -20.02 and 2.62. An expert who studies the effect on the elbow of varying amounts of force would then have to determine if this range of forces has any practical significance on injuries to the elbow of tennis players.

To illustrate that the separate-variance t test is less affected by unequal variances than is the pooled t test, the data from the computer simulation reported in Table 6.4 were analyzed using the separate-variance t test. The proportion of the 1,000 tests that incorrectly rejected H_0 is presented in Table 6.6. If the separate-variance t test was unaffected by the unequal variances, we would expect the proportions to be close to .05, the intended level, in all cases.

From the results in Table 6.6, we can observe that the separate-variance t test has a Type I error rate that is consistently very close to .05 in all the cases considered. On the other hand, the pooled t test has Type I error rates very different from .05 when the sample sizes are unequal and we sample from populations having very different variances.

In this section, we developed pooled-variance t methods based on the requirement of independent random samples from normal populations with equal population

TABLE 6.6
The effect of unequal variances on the Type I error rates of the separate-variance t test

		$\sigma_1 = k\sigma_2$				
n_1	n_2	$k = .25$	.50	1	2	4
10	10	.055	.040	.056	.038	.052
10	20	.055	.044	.049	.059	.051
10	40	.049	.047	.043	.041	.055
15	15	.044	.041	.054	.055	.057
15	30	.052	.039	.051	.043	.052
15	45	.058	.042	.055	.050	.058

variances. For situations when the variances are not equal, we introduced the separate-variance t' statistic. Confidence intervals and hypothesis tests based on these procedures (t or t') need not give identical results. Standard computer packages often report the results of both t and t' tests. Which of these results should you use in your report?

If the sample sizes are equal and the population variances are equal, the separate-variance t test and the pooled t test give algebraically identical results; that is, the computed t equals the computed t'. Thus, why not always use t' in place of t when $n_1 = n_2$? The reason we would select t over t' is that the df for t are nearly always larger than the df for t', and, hence, the power of the t test is greater than the power of the t' test when the variances are equal. When the sample sizes and variances are very unequal, the results of the t and t' procedures may differ greatly. The evidence in such cases indicates that the separate-variance methods are somewhat more reliable and more conservative than the results of the pooled t methods. However, if the populations have both different means and different variances, an examination of just the size of the difference in their means, $\mu_1 - \mu_2$, would be an inadequate description of how the populations differ. We should always examine the size of the differences in both the means and the standard deviations of the populations being compared. In Chapter 7, we will discuss procedures for examining the difference in the standard deviations of two populations.

6.3 A Nonparametric Alternative: The Wilcoxon Rank Sum Test

The two-sample t test of the previous section was based on several conditions: independent samples, normality, and equal variances. When the conditions of normality and equal variances are not valid but the sample sizes are large, the results using a t (or t') test are approximately correct. There is, however, an alternative test procedure that requires less stringent conditions. This procedure, called the **Wilcoxon rank sum test,** is discussed here.

Wilcoxon rank sum test

The assumptions for this test are that we have two independent random samples of sizes n_1 and n_2:

$$x_1, x_2, \ldots, x_{n_1} \quad \text{and} \quad y_1, y_2, \ldots, y_{n_2}$$

The population distributions of the xs and ys are identical with the exception that one distribution may be shifted to the right of the other distribution, as shown in Figure 6.4. We model this relationship by stating

$$y \stackrel{\mathrm{d}}{=} x + \Delta$$

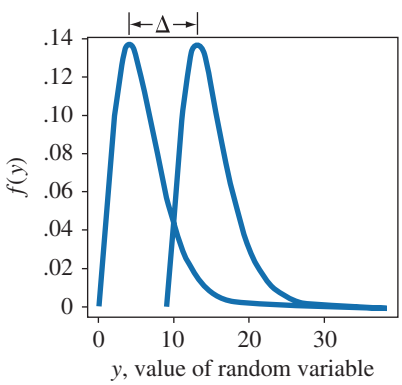


FIGURE 6.4 Skewed population distributions identical in shape but shifted

that the distribution of y equals the distribution of x plus a shift of size Δ. When Δ is a positive number, the population (treatment) associated with the y-values tends to have larger values than the population (treatment) associated with the x-values. In the previous section, $\Delta = \mu_1 - \mu_2$; that is, we were evaluating the difference in the population means. In this section, we will consider the difference in the populations more generally. Furthermore, the t-based procedures from Chapter 5 and Section 6.2 required that the population distributions have a normal distribution. The Wilcoxon rank sum test does not impose this restriction. Thus, the Wilcoxon procedure is more broadly applicable than the t-based procedures, especially for small sample sizes.

Because we are now allowing the population distributions to be nonnormal, the rank sum procedure must deal with the possibility of extreme observations in the data. One way to handle samples containing extreme values is to replace each data value with its rank (from lowest to highest) in the combined sample—that is, the sample consisting of the data from both populations. The smallest value in the combined sample is assigned the rank of 1, and the largest value is assigned the rank of $N = n_1 + n_2$. The ranks are not affected by how far the smallest (largest) data value is from next smallest (largest) data value. Thus, extreme values in data sets do not have as strong an effect on the rank sum statistic as they did in the t-based procedures.

The calculation of the rank sum statistic consists of the following steps:

1. List the data values in the combined data set from smallest to largest.
2. In the next column, assign the numbers 1 to N to the data values with 1 assigned to the smallest value and N to the largest value. These are the **ranks** of the observations.
3. If there are ties—that is, duplicated values—in the combined data set, the ranks for the observations in a tie are taken to be the average of the ranks for those observations.
4. Let T denote the sum of the ranks for the observations from population 1.

ranks

If the null hypothesis of identical population distributions is true, the n_1 ranks from population 1 are just a random sample from the N integers $1, \ldots, N$. Thus, under the null hypothesis, the distribution of the sums of the ranks T depends only on the sample sizes, n_1 and n_2, and does not depend on the shape of the population distributions. Under the null hypothesis, the sampling distribution of T has a mean and variance given by

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \quad \text{and} \quad \sigma_T^2 = \frac{n_1 n_2}{12}(n_1 + n_2 + 1)$$

Intuitively, if T is much smaller (or larger) than μ_T, we have evidence that the null hypothesis is false and in fact the population distributions are not equal. The rejection region for the rank sum test specifies the size of the difference between T and μ_T for the null hypothesis to be rejected. Because the distribution of T under the null hypothesis does not depend on the shape of the population distributions, Table 5 in the Appendix provides the critical values for the test regardless of the shape of the population distribution. The Wilcoxon rank sum test is summarized here.

Wilcoxon Rank Sum Test*

$(n_1 \leq 10, n_2 \leq 10)$

H_0: The two populations are identical ($\Delta = 0$).

H_a: **1.** Population 1 is shifted to the right of population 2 ($\Delta > 0$).
2. Population 1 is shifted to the left of population 2 ($\Delta < 0$).
3. Populations 1 and 2 are shifted from each other ($\Delta \neq 0$).

T.S.: T, the sum of the ranks in sample 1

R.R.: Use Table 5 in the Appendix to find critical values for T_U and T_L;
1. Reject H_0 if $T > T_U$ (one-tailed from Table 5).
2. Reject H_0 if $T < T_L$ (one-tailed from Table 5).
3. Reject H_0 if $T > T_U$ or $T < T_L$ (two-tailed from Table 5).

Check assumptions and draw conclusions.

*This test is equivalent to the Mann-Whitney U test (Conover, 1999).

After the completion of the test of hypotheses, we need to assess the size of the difference in the two populations (treatments). That is, we need to obtain a sample estimate of Δ and place a confidence interval on Δ. We use the Wilcoxon rank sum statistics to produce the confidence interval for Δ. First, obtain the $M = n_1 n_2$ possible differences in the two data sets: $x_i - y_j$ for $i = 1, \ldots, n_1$ and $j = 1 \ldots, n_2$. The estimator of Δ is the median of these M differences:

$$\hat{\Delta} = \text{median}[(x_i - y_j), \text{where } i = 1, \ldots, n_1 \text{ and } j = 1, \ldots, n_2]$$

Let $D_{(1)} \leq D_{(2)} \leq D_{(M)}$ denote the ordered values of the M differences, $x_i - y_j$. If $M = n_1 n_2$ is odd, take

$$\hat{\Delta} = D_{((M+1)/2)}$$

If $M = n_1 n_2$ is even, take

$$\hat{\Delta} = \frac{1}{2}[D_{(M/2)} + D_{(M/2+1)}]$$

We obtain a 95% confidence interval for Δ using the values from Table 5 in the Appendix for the Wilcoxon rank sum statistic. Let T_U be the $\alpha = .025$ one-tailed value from Table 5 in the Appendix, and let

$$C_{.025} = \frac{n_1(2n_2 + n_1 + 1)}{2} + 1 - T_U$$

If $C_{.025}$ is not an integer, take the nearest integer less than or equal to $C_{.025}$. The approximate 95% confidence interval for Δ, (Δ_L, Δ_U) is given by

$$\Delta_L = D_{(C_{.025})} \text{ and } \Delta_U = D_{(M+1-C_{.025})}$$

where $D_{(C_{.025})}$ and $D_{(M+1-C_{.025})}$ are obtained from the ordered values of all possible differences in the xs and ys.

For large values of n_1 and n_2, the value of $C_{\alpha/2}$ can be approximated using

$$C_{\alpha/2} = \frac{n_1 n_2}{2} - z_{\alpha/2}\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

where $z_{\alpha/2}$ is the percentile from the standard normal tables. We will illustrate these procedures in the following example.

EXAMPLE 6.5

Many states are considering lowering the blood-alcohol level at which a driver is designated as driving under the influence (DUI) of alcohol. An investigator for a legislative committee designed the following test to study the effect of alcohol on reaction time. Ten participants consumed a specified amount of alcohol. Another group of 10 participants consumed the same amount of a nonalcoholic drink, a placebo. The two groups did not know whether they were receiving alcohol or the placebo. The 20 participants' average reaction times (in seconds) to a series of simulated driving situations are reported in Table 6.7. Does it appear that alcohol consumption increases reaction time?

TABLE 6.7
Data for Example 6.5

Placebo	0.90	0.37	1.63	0.83	0.95	0.78	0.86	0.61	0.38	1.97
Alcohol	1.46	1.45	1.76	1.44	1.11	3.07	0.98	1.27	2.56	1.32

a. Why is the t test inappropriate for analyzing the data in this study?

b. Use the Wilcoxon rank sum test to test the hypotheses:

H_0: The distributions of reaction times for the placebo and alcohol populations are identical ($\Delta = 0$).

H_a: The distribution of reaction times for the placebo consumption population is shifted to the left of the distribution for the alcohol population. (Larger reaction times are associated with the consumption of alcohol, $\Delta < 0$.)

c. Place 95% confidence intervals on the median reaction times for the two groups and on Δ.

d. Compare the results you obtain to the results from a software program.

Solution

a. A boxplot of the two samples is given in Figure 6.5. The plots indicate that the population distributions are skewed to the right because 10% of the data values are large outliers and the upper whiskers are longer than the lower whiskers. The sample sizes are both small, and, hence, the t test may be inappropriate for analyzing this study.

b. The Wilcoxon rank sum test will be conducted to evaluate whether alcohol consumption increases reaction time. Table 6.8 contains the ordered data for the combined samples, along with their associated ranks. We will designate observations from the placebo group as 1 and from the alcohol group as 2.

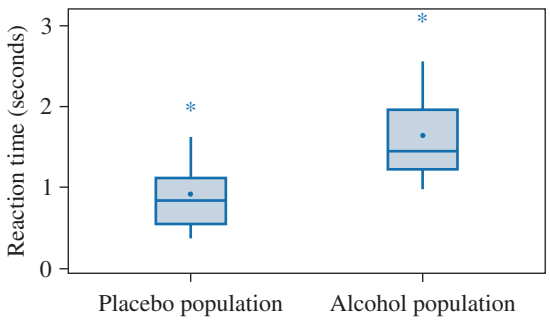


FIGURE 6.5
Boxplots of placebo and alcohol populations (means are indicated by solid circles)

TABLE 6.8
Ordered reaction times and ranks

	Ordered Data	Group	Rank		Ordered Data	Group	Rank
1	0.37	1	1	11	1.27	2	11
2	0.38	1	2	12	1.32	2	12
3	0.61	1	3	13	1.44	2	13
4	0.78	1	4	14	1.45	2	14
5	0.83	1	5	15	1.46	2	15
6	0.86	1	6	16	1.63	1	16
7	0.90	1	7	17	1.76	2	17
8	0.95	1	8	18	1.97	1	18
9	0.98	2	9	19	2.56	2	19
10	1.11	2	10	20	3.07	2	20

For $\alpha = .05$, reject H_0 if $T < 83$, using Table 5 in the Appendix with $\alpha = .05$, one-tailed, and $n_1 = n_2 = 10$. The value of T is computed by summing the ranks from group 1: $T = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 16 + 18 = 70$. Because 70 is less than 83, we reject H_0 and conclude there is significant evidence that the placebo population has smaller reaction times than the population of alcohol consumers.

c. Because we have small sample sizes and the population distributions appear to be skewed to the right, we will construct confidence intervals on the median reaction times in place of confidence intervals on the mean reaction times. Using the methodology from Section 5.9 and Table 4 in the Appendix, we find

$$C_{\alpha(2),n} = C_{.05,10} = 1$$

Thus,

$$L_{.025} = C_{.05,10} + 1 = 2$$

and

$$U_{.025} = n - C_{.05,10} = 10 - 1 = 9$$

The 95% confidence intervals for the population medians are given by

$$(M_L, M_U) = (y_{(2)}, y_{(9)})$$

Thus, a 95% confidence interval is (.38, 1.63) for the placebo population median and (1.11, 2.56) for the alcohol population median. Because the sample sizes are very small, the confidence intervals are not very informative.

To compute the 95% confidence interval for Δ, we need to form the $M = n_1 n_2 = 10(10) = 100$ possible differences $D_{ij} = y_{1i} - y_{2j}$. Next, we obtain the $\alpha = .025$ value of T_U from Table 5 in the Appendix with $n_1 = n_2 = 10$—that is, $T_U = 131$. Using the formula for $C_{.025}$, we obtain

$$C_{.025} = \frac{n_1(2n_2 + n_1 + 1)}{2} + 1 - T_U = \frac{10(2(10) + 10 + 1)}{2} + 1 - 131 = 25$$

$$\Delta_L = D_{(C_{.025})} = D_{(25)} \text{ and } \Delta_U = D_{(M+1-C_{.025})} = D_{(100+1-25)} = D_{(76)}$$

Thus, we need to find the 25th and 76th ordered values of the differences $D_{ij} = x_i - y_j$. Table 6.9 contains the 100 differences, Ds. We would next sort the Ds from smallest to largest. The estimator of Δ would be the median of the differences:

$$\hat{\Delta} = \frac{1}{2}[D_{(50)} + D_{(51)}] = \frac{1}{2}[(-0.61) + (-0.61)] = -0.61$$

To obtain an approximate 95% confidence interval for Δ, we first need to obtain

$$D_{(25)} = -1.07 \quad \text{and} \quad D_{(76)} = -0.28$$

Therefore, our approximate 95% confidence interval for Δ is $(-1.07, -0.28)$.

d. The output from Minitab is given here.

```
Mann-Whitney Confidence Interval and Test

PLACEBO  N =  10     Median =       0.845
ALCOHOL  N =  10     Median =       1.445
Point estimate for ETA1-ETA2 is    −0.610
95.5 Percent CI for ETA1-ETA2 is (−1.080,−0.250)
W = 70.0
Test of ETA1 = ETA2 vs ETA1 < ETA2 is significant at 0.0046
```

TABLE 6.9 Summary data for Example 6.5

y_{1i}	y_{2j}	D_{ij}	y_{1i}	y_{2j}	D_{ij}	y_{1i}	y_{2j}	D_{ij}	y_{1i}	y_{2j}	D_{ij}	y_{1i}	y_{2j}	D_{ij}
.90	1.46	−.56	.37	1.46	−1.09	1.63	1.46	.17	.83	1.46	−.63	.95	1.46	−.51
.90	1.45	−.55	.37	1.45	−1.08	1.63	1.45	.18	.83	1.45	−.62	.95	1.45	−.50
.90	1.76	−.86	.37	1.76	−1.39	1.63	1.76	−.13	.83	1.76	−.93	.95	1.76	−.81
.90	1.44	−.54	.37	1.44	−1.07	1.63	1.44	.19	.83	1.44	−.61	.95	1.44	−.49
.90	1.11	−.21	.37	1.11	−.74	1.63	1.11	.52	.83	1.11	−.28	.95	1.11	−.16
.90	3.07	−2.17	.37	3.07	−2.70	1.63	3.07	−1.44	.83	3.07	−2.24	.95	3.07	−2.12
.90	0.98	−.08	.37	.98	−.61	1.63	.98	.65	.83	.98	−.15	.95	.98	−.03
.90	1.27	−.37	.37	1.27	−.90	1.63	1.27	.36	.83	1.27	−.44	.95	1.27	−.32
.90	2.56	−1.66	.37	2.56	−2.19	1.63	2.56	−.93	.83	2.56	−1.73	.95	2.56	−1.61
.90	1.32	−.42	.37	1.32	−.95	1.63	1.32	.31	.83	1.32	−.49	.95	1.32	−.37
.78	1.46	−.68	.86	1.46	−.60	.61	1.46	−.85	.38	1.46	−1.08	1.97	1.46	.51
.78	1.45	−.67	.86	1.45	−.59	.61	1.45	−.84	.38	1.45	−1.07	1.97	1.45	.52
.78	1.76	−.98	.86	1.76	−.90	.61	1.76	−1.15	.38	1.76	−1.38	1.97	1.76	.21
.78	1.44	−.66	.86	1.44	−.58	.61	1.44	−.83	.38	1.44	−1.06	1.97	1.44	.53
.78	1.11	−.33	.86	1.11	−.25	.61	1.11	−.50	.38	1.11	−.73	1.97	1.11	.86
.78	3.07	−2.29	.86	3.07	−2.21	.61	3.07	−2.46	.38	3.07	−2.69	1.97	3.07	−1.10
.78	.98	−.20	.86	.98	−.12	.61	.98	−.37	.38	.98	−.60	1.97	.98	.99
.78	1.27	−.49	.86	1.27	−.41	.61	1.27	−.66	.38	1.27	−.89	1.97	1.27	.70
.78	2.56	−1.78	.86	2.56	−1.70	.61	2.56	−1.95	.38	2.56	−2.18	1.97	2.56	−.59
.78	1.32	−.54	.86	1.32	−.46	.61	1.32	−.71	.38	1.32	−.94	1.97	1.32	.65

Minitab refers to the test statistic as the Mann-Whitney test. This test is equivalent to the Wilcoxon test statistic. In fact, the value of the test statistic $W = 70$ is identical to the Wilcoxon $T = 70$. The output indicates that the p-value $= .0046$ and a 95.5% confidence interval for Δ is given by $(-1.08, -.25)$.

Note: This interval is slightly different from the interval computed in part (c) because Minitab computed a 95.6% confidence interval, whereas we computed a 94.8% confidence interval. ■

When both sample sizes are more than 10, the sampling distribution of T is approximately normal; this allows us to use a z statistic in place of T when using the Wilcoxon rank sum test:

$$z = \frac{T - \mu_T}{\sigma_T}$$

The theory behind the Wilcoxon rank sum test requires that the population distributions be continuous, so the probability that any two data values are equal is zero. Because in most studies we record data values to only a few decimal places, we will often have ties—that is, observations with the same value. For these situations, each observation in a set of tied values receives a rank score equal to the average of the ranks for the set of values. When there are ties, the variance of T must be adjusted. The adjusted value of σ_T^2 is shown here.

$$\sigma_T^2 = \frac{n_1 n_2}{12}\left((n_1 + n_2 + 1) - \frac{\sum_{j=1}^{k} t_j(t_j^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}\right)$$

where k is the number of tied groups and t_j denotes the number of tied observations in the jth group. Note that when there are no tied observations, $t_j = 1$ for all j, which results in

$$\sigma_T^2 = \frac{n_1 n_2}{12}(n_1 + n_2 + 1)$$

From a practical standpoint, unless there are many ties, the adjustment will result in very little change to σ_T^2. The normal approximation to the Wilcoxon rank sum test is summarized here.

Wilcoxon Rank Sum Test: Normal Approximation

$n_1 > 10$ and $n_2 > 10$

H_0: The two populations are identical.

H_a:
1. Population 1 is shifted to the right of population 2.
2. Population 1 is shifted to the left of population 2.
3. Population 1 and 2 are shifted from each other.

T.S.: $z = \dfrac{T - \mu_T}{\sigma_T}$, where T denotes the sum of the ranks in sample 1

R.R.: For a specified value of α,
1. Reject H_0 if $z \geq z_\alpha$.
2. Reject H_0 if $z \leq -z_\alpha$.
3. Reject H_0 if $|z| \geq z_{\alpha/2}$.

Check assumptions and draw conclusions.

EXAMPLE 6.6

Environmental engineers were interested in determining whether a cleanup project on a nearby lake was effective. Prior to initiation of the project, they obtained 12 water samples at random from the lake and analyzed the samples for the amount of dissolved oxygen (in ppm). Due to diurnal fluctuations in the dissolved oxygen, all measurements were obtained at the 2 P.M. peak period. The before and after data are presented in Table 6.10.

TABLE 6.10
Dissolved oxygen measurements (in ppm)

Before Cleanup		After Cleanup	
11.0	11.6	10.2	10.8
11.2	11.7	10.3	10.8
11.2	11.8	10.4	10.9
11.2	11.9	10.6	11.1
11.4	11.9	10.6	11.1
11.5	12.1	10.7	11.3

a. Use $\alpha = .05$ to test the following hypotheses:

H_0: The distributions of dissolved oxygen measurements taken before the cleanup project and 6 months after the cleanup project began are identical.

H_a: The distribution of dissolved oxygen measurements taken before the cleanup project is shifted to the right of the corresponding distribution of measurements taken 6 months after the cleanup project began. (Note that a cleanup project has been effective in one sense if the dissolved oxygen level drops over a period of time.)

For convenience, the data are arranged in ascending order in Table 6.10.

b. Has the correction for ties made much of a difference?

Solution

a. First, we must jointly rank the combined sample of 24 observations by assigning the rank of 1 to the smallest observation, the rank of 2 to the next smallest, and so on. When two or more measurements are the same, we assign all of them a rank equal to the average of the ranks they occupy. The sample measurements and associated ranks (shown in parentheses) are listed in Table 6.11.

Because n_1 and n_2 are both greater than 10, we will use the test statistic z. If we are trying to detect a shift to the left in the distribution after the cleanup, we expect the sum of the ranks for the observations in sample 1 to be large. Thus, we will reject H_0 for large values of $z = (T - \mu_T)/\sigma_T$.

Grouping the measurements with tied ranks, we have 18 groups. These groups are listed in Table 6.12 with the corresponding values of t_j, the number of tied ranks in the group.

For all groups with $t_j = 1$, there is no contribution for

$$\frac{\Sigma_j t_j(t_j^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}$$

in σ_T^2 because $t_j^2 - 1 = 0$. Thus, we will need only $t_j = 2, 3$.

TABLE 6.11
Dissolved oxygen measurements and ranks

Before Cleanup		After Cleanup	
11.0	(10)	10.2	(1)
11.2	(14)	10.3	(2)
11.2	(14)	10.4	(3)
11.2	(14)	10.6	(4.5)
11.4	(17)	10.6	(4.5)
11.5	(18)	10.7	(6)
11.6	(19)	10.8	(7.5)
11.7	(20)	10.8	(7.5)
11.8	(21)	10.9	(9)
11.9	(22.5)	11.1	(11.5)
11.9	(22.5)	11.1	(11.5)
12.1	(24)	11.3	(16)
	$T = 216$		

TABLE 6.12
Ranks, groups, and ties

Rank	Group	t_j	Rank	Group	t_j
1	1	1	14, 14, 14	10	3
2	2	1	16	11	1
3	3	1	17	12	1
4.5, 4.5	4	2	18	13	1
6	5	1	19	14	1
7.5, 7.5	6	2	20	15	1
9	7	1	21	16	1
10	8	1	22.5, 22.5	17	2
11.5, 11.5	9	2	24	18	1

Substituting our data in the formulas, we obtain

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{12(12 + 12 + 1)}{2} = 150$$

$$\sigma_T^2 = \frac{n_1 n_2}{12}\left[(n_1 + n_2 + 1) - \frac{\Sigma t_j(t_j^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}\right]$$

$$= \frac{12(12)}{12}\left[25 - \frac{6 + 6 + 6 + 24 + 6}{24(23)}\right]$$

$$= 12(25 - .0870) = 298.956$$

$$\sigma_T = 17.29$$

The computed value of z is

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{216 - 150}{17.29} = 3.82$$

Using the R function **pnorm**(z_c), the test statistic $z = 3.82$ has p-value $P(z \geq 3.82) = 1 - \textbf{pnorm}(3.82) = .00007$. This implies that there is very strong evidence in the data that the distribution of before-cleanup measurements is shifted to the right of the corresponding distribution of after-cleanup measurements; that is, the after-cleanup measurements of dissolved oxygen tend to be smaller than the corresponding before-cleanup measurements.

b. The value of σ_T^2 without correcting for ties is

$$\sigma_T^2 = \frac{12(12)(25)}{12} = 300 \quad \text{and} \quad \sigma_T = 17.32$$

For this value of σ_T, $z = 3.81$ rather than 3.82, which was found by applying the correction. This should help you understand how little effect the correction has on the final result unless there are a large number of ties. ■

The Wilcoxon rank sum test is an alternative to the two-sample t test, with the rank sum test requiring fewer conditions than the t test. In particular, the rank sum test does not require the two populations to have normal distributions; it requires only that the distributions be identical except possibly that one distribution could be shifted from the other distribution. When both distributions are normal, the t test is more likely to detect an existing difference; that is, the t test has greater power than the rank sum test. This is logical because the t test uses the magnitudes of the observations rather than just their relative magnitudes (ranks), as is done in the rank sum test. However, when the two distributions are nonnormal, the Wilcoxon rank sum test has greater power; that is, it is more likely to detect a shift in the population distributions. Also, the level or probability of a Type I error for the Wilcoxon rank sum test will be equal to the stated level for all population distributions. The t test's *actual* level will deviate from its stated value when the population distributions are nonnormal. This is particularly true when nonnormality of the population distributions is present in the form of severe skewness or extreme outliers.

Randles and Wolfe (1979) investigated the effect of skewed and heavy-tailed distributions on the power of the t test and the Wilcoxon rank sum test. Table 6.13 contains a portion of the results of their simulation study. For each set of distributions, sample sizes and shifts in the populations, 5,000 samples were drawn, and the proportion of times a level $\alpha = .05$ t test or Wilcoxon rank sum test rejected H_0 was recorded. The distributions considered were normal, double exponential (symmetric, heavy-tailed), Cauchy (symmetric, extremely heavy-tailed), and Weibull (skewed to the right). Shifts of size 0, $.6\sigma$, and 1.2σ were considered, where σ denotes the standard deviation of the distribution, with the exception of the Cauchy distribution, where σ is a general scale parameter.

When the distribution is normal, the t test is only slightly better—has greater power values—than the Wilcoxon rank sum test. For the double exponential, the Wilcoxon test has greater power than the t test. For the Cauchy distribution, the

TABLE 6.13
Power of t test (t) and Wilcoxon rank sum test (T) with $\alpha = .05$

Distribution		Normal			Double Exponential			Cauchy			Weibull		
	Shift	0	.6	1.2	0	.6	1.2	0	.6	1.2	0	.6	1.2
n_1, n_2	Test												
5, 5	t	.044	.213	.523	.045	.255	.588	.024	.132	.288	.049	.221	.545
	T	.046	.208	.503	.049	.269	.589	.051	.218	.408	.049	.219	.537
5, 15	t	.047	.303	.724	.046	.304	.733	.056	.137	.282	.041	.289	.723
	T	.048	.287	.694	.047	.351	.768	.046	.284	.576	.049	.290	.688
15, 15	t	.052	.497	.947	.046	.507	.928	.030	.153	.333	.046	.488	.935
	T	.054	.479	.933	.046	.594	.962	.046	.484	.839	.046	.488	.927

level of the t test deviates significantly from .05, and its power is much lower than for the Wilcoxon test. When the distribution was somewhat skewed, as in the Weibull distribution, the tests had similar performance. Furthermore, the level and power of the t test were nearly identical to the values when the distribution was normal. The t test is quite robust to skewness except when there are numerous extreme values.

6.4 Inferences About $\mu_1 - \mu_2$: Paired Data

The methods we presented in the preceding three sections were appropriate for situations in which independent random samples are obtained from two populations. These methods are not appropriate for studies or experiments in which each measurement in one sample is *matched* or *paired* with a particular measurement in the other sample. In this section, we will deal with methods for analyzing "paired" data. We begin with an example.

EXAMPLE 6.7

Insurance adjusters are concerned about the high estimates they are receiving for auto repairs from garage I compared to garage II. To verify their suspicions, each of 15 cars recently involved in an accident was taken to both garages for separate estimates of repair costs. The estimates from the two garages are given in Table 6.14.

A preliminary analysis of the data used a two-sample t test.

Solution Computer output for these data is shown here.

```
Two-Sample T-Test and Confidence Interval

Two-sample T for Garage I vs Garage II

             N      Mean     StDev   SE Mean
Garage I    15     16.85      3.20      0.83
Garage II   15     16.23      2.94      0.76

95% CI for mu Garage I - mu Garage II: (-1.69, 2.92)
T-Test mu Garage I = mu Garage II (vs not =): T = 0.55  P = 0.59  DF = 27
```

TABLE 6.14
Repair estimates (in hundreds of dollars)

Car	Garage I	Garage II
1	17.6	17.3
2	20.2	19.1
3	19.5	18.4
4	11.3	11.5
5	13.0	12.7
6	16.3	15.8
7	15.3	14.9
8	16.2	15.3
9	12.2	12.0
10	14.8	14.2
11	21.3	21.0
12	22.1	21.0
13	16.9	16.1
14	17.6	16.7
15	18.4	17.5
Totals:	$\bar{y}_1 = 16.85$	$\bar{y}_2 = 16.23$
	$s_1 = 3.20$	$s_1 = 2.94$

From the output, we see there is a consistent difference in the sample means ($\bar{y}_1 - \bar{y}_2 = .62$). However, this difference is rather small considering the variability of the measurements ($s_1 = 3.20$, $s_2 = 2.94$). In fact, the computed t-value (.55) has a p-value of .59, indicating very little evidence of a difference in the average claim estimates for the two garages. ■

A closer glance at the data in Table 6.14 indicates that something about the conclusion in Example 6.7 is inconsistent with our intuition. For all but one of the 15 cars, the estimate from garage I was higher than that from garage II. From our knowledge of the binomial distribution, the probability of observing garage I estimates higher in $y = 14$ or more of the $n = 15$ trials, assuming no difference ($\pi = .5$) for garages I and II, is

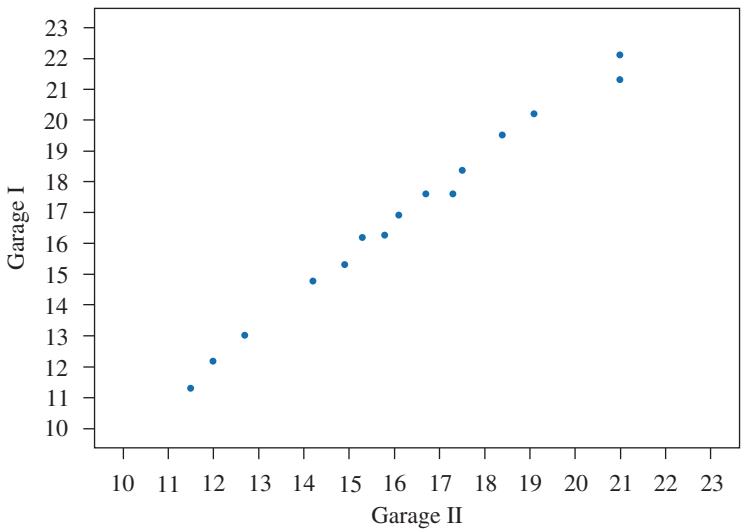
$$P(y = 14 \text{ or } 15) = P(y = 14) + P(y = 15)$$
$$= \binom{15}{14}(.5)^{14}(.5) + \binom{15}{15}(.5)^{15} = .000488$$

Thus, if the two garages in fact have the same distribution of estimates, there is approximately a 5 in 10,000 chance of having 14 or more estimates from garage I higher than those from garage II. Using this probability, we would argue that the observed estimates are highly contradictory to the null hypothesis of equality of distribution of estimates for the two garages. Why are there such conflicting results from the t test and the binomial calculation?

The explanation of the difference in the conclusions from the two procedures is that one of the required conditions for the t test, two samples being independent of each other, has been violated by the manner in which the study was conducted. The adjusters obtained a measurement from both garages for each car. For the two samples to be independent, the adjusters would have to take a random sample of 15 cars to garage I and a *different* random sample of 15 to garage II.

As can be observed in Figure 6.6, the repair estimates for a given car are about the same value, but there is a large variability in the estimates from each garage. The large variability *among* the 15 estimates from each garage diminishes the relative size of any difference *between* the two garages. When designing the study, the adjusters recognized that the large differences in the amount of damage

FIGURE 6.6
Repair estimates from two garages



suffered by the cars would result in a large variability in the 15 estimates at both garages. By having both garages give an estimate on each car, the adjusters could calculate the difference between the estimates from the garages and hence reduce the large car-to-car variability.

This example illustrates a general design principle. In many situations, the available experimental units may be considerably different prior to their random assignment to the treatments with respect to characteristics that may affect the experimental responses. These differences will often then mask true treatment differences. In the previous example, the cars had large differences in the amount of damage suffered during the accident and hence would be expected to have large differences in their repair estimates no matter what garage gave the repair estimate. When comparing two treatments or groups in which the available experimental units have important differences prior to their assignment to the treatments or groups, the samples should be paired. There are many ways to design experiments to yield paired data. One method involves having the same group of experimental units receive both treatments, as was done in the repair estimates example. A second method involves having measurements taken before and after the treatment is applied to the experimental units. For example, suppose we want to study the effect of a new medicine proposed to reduce blood pressure. We would record the blood pressure of participants before they received the medicine and then after receiving the medicine. A third design procedure uses naturally occurring pairs such as twins or spouses. A final method pairs the experimental units with respect to factors that may mask differences in the treatments. For example, a study is proposed to evaluate two methods for teaching remedial reading. The participants could be paired based on a pretest of their reading ability. After pairing the participants, the two methods are randomly assigned to the participants within each pair.

A proper analysis of paired data needs to take into account the lack of independence between the two samples. The sampling distribution for the difference in the sample means, $\bar{y}_1 - \bar{y}_2$, will have a mean and standard error of

$$\mu_{\bar{y}_1-\bar{y}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma_{\bar{y}_1-\bar{y}_2} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho}{n}}$$

where ρ measures the amount of dependence between the two samples. When the two samples produce similar measurements, ρ is positive and the standard error of $\bar{y}_1 - \bar{y}_2$ is smaller than what would be obtained using two independent samples. This was the case in the repair estimates data. The size and sign of ρ can be determined by examining the plot of the paired data values. The magnitude of ρ is large when the plotted points are close to a straight line. The sign of ρ is positive when the plotted points follow an increasing line and negative when the plotted points follow a decreasing line. From Figure 6.6, we observe that the estimates are close to an increasing line, and, thus, ρ will be positive. Using paired data in the repair estimate study will reduce the variability in the standard error of the difference in the sample means in comparison to using independent samples.

The actual analysis of paired data requires us to compute the differences in the n pairs of measurements, $d_i = y_{1i} - y_{2i}$, and obtain $\bar{d}$, s_d, and the mean and standard deviations in the d_is. Also, we must transform the hypotheses about μ_1 and μ_2 into hypotheses about the mean of the differences, $\mu_d = \mu_1 - \mu_2$. The conditions required to develop a t procedure for testing hypotheses and constructing confidence intervals for μ_d are

1. The sampling distribution of the d_is is a normal distribution.
2. The d_is are independent; that is, the pairs of observations are independent.

A summary of the test procedure is given here.

Paired *t* test

H_0: **1.** $\mu_d \le D_0$ (D_0 is a specified value, often .0)
2. $\mu_d \ge D_0$
3. $\mu_d = D_0$

H_a: **1.** $\mu_d > D_0$
2. $\mu_d < D_0$
3. $\mu_d \ne D_0$

T.S.: $t = \dfrac{\bar{d} - D_0}{s_d/\sqrt{n}}$

R.R.: For a level α, Type I error rate with df $= n - 1$
1. Reject H_0 if $t \ge t_\alpha$.
2. Reject H_0 if $t \le -t_\alpha$.
3. Reject H_0 if $|t| \ge t_{\alpha/2}$.

Check assumptions and draw conclusions.

The corresponding $100(1 - \alpha)\%$ confidence interval on $\mu_d = \mu_1 - \mu_2$ based on the paired data is shown here.

$100(1 - \alpha)\%$ Confidence Interval for μ_d Based on Paired Data

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

where n is the number of pairs of observations (and hence the number of differences) and df $= n - 1$.

EXAMPLE 6.8

Refer to the data of Example 6.7, and perform a paired t test. Draw a conclusion based on $\alpha = .05$.

Solution For these data, the parts of the statistical test are

H_0: $\mu_d = \mu_1 - \mu_2 \le 0$

H_a: $\mu_d > 0$

T.S.: $t = \dfrac{\bar{d}}{s_d/\sqrt{n}}$

R.R.: For df $= n - 1 = 14$, reject H_0 if $t \ge t_{.05}$.

Before computing t, we must first calculate $\bar{d}$ and s_d . For the data of Table 6.14, we have the differences d_i = garage I estimate − garage II estimate (see Table 6.15).

TABLE 6.15 Difference data from Table 6.14

Car	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
d_i	.3	1.1	1.1	−.2	.3	.5	.4	.9	.2	.6	.3	1.1	.8	.9	.9

The mean and standard deviation are given here.

$$\bar{d} = .61 \quad \text{and} \quad s_d = .394$$

Substituting into the test statistic t, we have

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{.61}{.394/\sqrt{15}} = 6.00$$

Indeed, $t = 6.00$ is far beyond all tabulated t values for df = 14, so the p-value is less than .005; in fact, the p-value is .000016. We conclude that the mean repair estimate for garage I is greater than that for garage II. This conclusion agrees with our intuitive finding based on the binomial distribution.

The point of all this discussion is not to suggest that we typically have two or more analyses that may give *very* conflicting results for a given situation. Rather, the point is that the analysis must fit the experimental situation. For this experiment, the samples are dependent, demanding that we use an analysis appropriate for dependent (paired) data.

After determining that there is a *statistically significant* difference in the means, we should estimate the size of the difference. A 95% confidence interval for $\mu_1 - \mu_2 = \mu_d$ will provide an estimate of the size of the difference in the average repair estimate between the two garages:

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

$$.61 \pm 2.145\frac{.394}{\sqrt{15}} = .61 \pm .22 = (.39, .83)$$

Thus, we are 95% confident that the mean repair estimates differ by a value between \$390 and \$830. The insurance adjusters determined that a difference of this size is of practical significance. ■

Reducing the standard error of $\bar{y}_1 - \bar{y}_2$ by using the differences, d_is, in place of the observed values, y_{1i}s and y_{2i}s, will often produce a t test having greater power and confidence intervals having smaller width. Is there any loss in using paired data experiments? Yes, the t procedures using the d_is have df $= n - 1$, whereas the t procedures using the individual measurements have df $= n_1 + n_2 - 2 = 2(n - 1)$. Thus, when designing a study or experiment, the choice between using an independent samples experiment and a paired data experiment will depend on how much difference exists in the experimental units prior to their assignment to the treatments. If there are only small differences, then the independent samples design is more efficient. If the differences in the experimental units are extreme, then the paired data design is more efficient, provided that the two measurements within the pairs are positively correlated.

6.5 A Nonparametric Alternative: The Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test, which makes use of the sign and the magnitude of the rank of the differences between pairs of measurements, provides an alternative to the paired t test when the population distribution of the differences is nonnormal. The Wilcoxon signed-rank test requires that the population distribution of differences be symmetric about the unknown median M. Let D_0 be a specified

hypothesized value of M. The test evaluates shifts in the distribution of differences to the right or left of D_0; in most cases, D_0 is 0. The computation of the signed-rank test involves the following steps:

1. Calculate the differences in the n pairs of observations.
2. Subtract D_0 from all the differences.
3. Delete all zero values. Let n be the number of nonzero values.
4. List the *absolute values* of the differences in increasing order, and assign them the ranks $1, \ldots, n$ (or the average of the ranks for ties).

We define the following notation before describing the Wilcoxon signed-rank test:

n = the number of pairs of observations with a nonzero difference
T_+ = the sum of the positive ranks; if there are no positive ranks, $T = 0$
T_- = the sum of the negative ranks; if there are no negative ranks, $T = 0$
T = the smaller of T_+ and T_-

μ_T

$$\mu_T = \frac{n(n+1)}{4}$$

σ_T

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

g groups

If we group together all differences assigned the same rank and there are g such **groups**, the variance of T is

$$\sigma_T^2 = \frac{1}{24}\left[n(n+1)(2n+1) - \frac{1}{2}\sum_j t_j(t_j-1)(t_j+1)\right]$$

t_j

where t_j is the number of tied ranks in the jth group. Note that if there are no tied ranks, $t_j = 1$ for all groups. The formula then reduces to

$$\sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$$

The Wilcoxon signed-rank test is presented here. Let M be the median of the population of differences.

Wilcoxon Signed-Rank Test

H_0: $M = D_0$ (D_0 is specified; generally D_0 is set to 0.)

H_a:
1. $M > D_0$
2. $M < D_0$
3. $M \neq D_0$

($n \leq 50$)

T.S.:
1. $T = T_-$
2. $T = T_+$
3. T = smaller of T_+ and T_-

R.R.: For a specified value of α (one-tailed .05, .025, .01, or .005; two-tailed .10, .05, .02, .01) and fixed number of nonzero differences n, reject H_0 if the value of T is less than or equal to the appropriate entry in Table 6 in the Appendix.

($n > 50$)

T.S.: Compute the test statistic

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

R.R.: For cases 1 and 2, reject H_0 if $z < -z_\alpha$; for case 3, reject H_0 if $z < -z_{\alpha/2}$.

Check assumptions, place a confidence interval on the median of the differences, and state conclusions.

EXAMPLE 6.9

A city park department compared a new formulation of a fertilizer, brand A, to the previously used fertilizer, brand B, on each of 20 different softball fields. Each field was divided in half, with brand A randomly assigned to one half of the field and brand B to the other. Sixty pounds of fertilizer per acre were then applied to the fields. The effect of the fertilizer on the grass grown at each field was measured by the weight (in pounds) of grass clippings produced by mowing the grass at the fields over a 1-month period. Evaluate whether brand A tends to produce more grass than brand B. The data are given in Table 6.16.

TABLE 6.16

Field	Brand A	Brand B	Difference	Field	Brand A	Brand B	Difference
1	211.4	186.3	25.1	11	208.9	183.6	25.3
2	204.4	205.7	−1.3	12	208.7	188.7	20.0
3	202.0	184.4	17.6	13	213.8	188.6	25.2
4	201.9	203.6	−1.7	14	201.6	204.2	−2.6
5	202.4	180.4	22.0	15	201.8	181.6	20.1
6	202.0	202.0	0	16	200.3	208.7	−8.4
7	202.4	181.5	20.9	17	201.8	181.5	20.3
8	207.1	186.7	20.4	18	201.5	208.7	−7.2
9	203.6	205.7	−2.1	19	212.1	186.8	25.3
10	216.0	189.1	26.9	20	203.4	182.9	20.5

Solution Evaluate whether brand A tends to produce more grass than brand B. Plots of the differences in grass yields for the 20 fields are given in Figures 6.7(a) and (b). The differences appear to not follow a normal distribution and appear to form two distinct clusters. Thus, we will apply the Wilcoxon signed-rank test to evaluate the differences in grass yields from brand A and brand B. The null hypothesis is that the distribution of differences is symmetrical about 0 against the alternative that the differences tend to be greater than 0. First, we must rank (from smallest to largest) the absolute values of the $n = 20 - 1 = 19$ nonzero differences. These ranks appear in Table 6.17.

FIGURE 6.7(a)
Boxplot of differences (with H_0 and 95% t confidence interval for the mean)

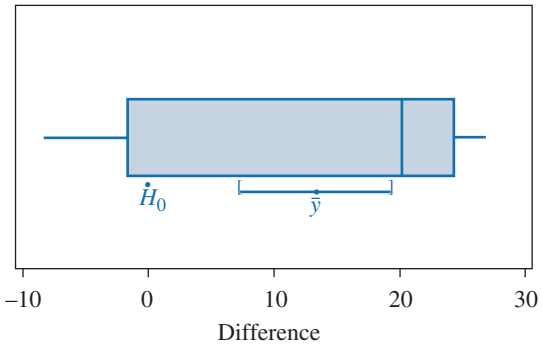


FIGURE 6.7(b)
Normal probability plot of differences

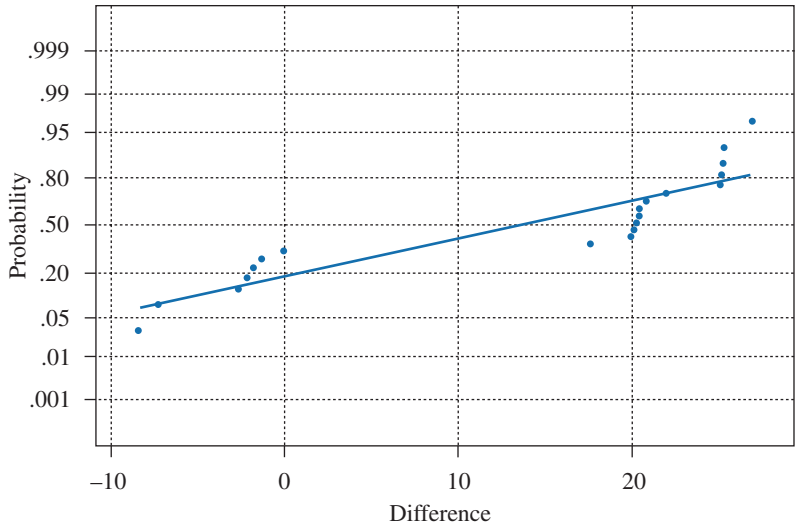


TABLE 6.17
Rankings of grass yield data

Field	Difference	Rank of Absolute Difference	Sign of Difference	Field	Difference	Rank of Absolute Difference	Sign of Difference
1	25.1	15	Positive	11	25.3	17.5	Positive
2	−1.3	1	Negative	12	20.0	8	Positive
3	17.6	7	Positive	13	25.2	16	Positive
4	−1.7	2	Negative	14	−2.6	4	Negative
5	22.0	14	Positive	15	20.1	9	Positive
6	0	None	Positive	16	−8.4	6	Negative
7	20.9	13	Positive	17	20.3	10	Positive
8	20.4	11	Positive	18	−7.2	5	Negative
9	−2.1	3	Negative	19	25.3	17.5	Positive
10	26.9	19	Positive	20	20.5	12	Positive

The sums of the positive and negative ranks are

$$T_- = 1 + 2 + 3 + 4 + 5 + 6 = 21$$

and

$$T_+ = 7 + 8 + 9 + 10 + 11 + 12 + 13 + 14 + 15 + 16 + 17.5 + 17.5 + 19 = 169$$

Because H_a: $M > 0$, $T = T_- = 21$. For a one-sided test with $n = 19$ and $\alpha = .05$, we see from Table 6 in the Appendix that we will reject H_0 if T is less than or equal to 53. Thus, we reject H_0 and conclude that brand A fertilizer tends to produce more grass than does brand B.

A 95% confidence interval on the median difference in grass production is obtained by using the methods given in Chapter 5. Because the number of sample differences is an even number, the estimated median difference is obtained by taking the average of the 10th- and 11th-largest differences: $D_{(10)}$ and $D_{(11)}$:

$$\hat{M} = \frac{1}{2}[D_{(10)} + D_{(11)}] = \frac{1}{2}[20.1 + 20.3] = 20.2$$

A 95% confidence interval for M is obtained as follows. From Table 4 in the Appendix with $\alpha(2) = .05$, we have $C_{\alpha(2),\,20} = 5$. Therefore,

$$L_{.025} = C_{.05,\,20} = 5 + 1 = 6$$

and

$$U_{.025} = n - C_{.05,\,20} = 20 - 5 = 15$$

The 95% confidence for the median of population of differences is

$$(M_L, M_U) = (D_6, D_{15}) = (-1.3, 22.0) \blacksquare$$

The choice of an appropriate paired-sample test depends on examining different types of deviations from normality. Because the level of the Wilcoxon signed-rank does not depend on the population distribution, it is the same as the stated value for all symmetric distributions. The level of the paired t test may be different from its stated value when the population distribution is very nonnormal. Also, we need to examine which test has greater power. We will report a portion of a simulation study contained in ***Randles and Wolfe (1979)***. The population distributions considered were normal, uniform (short-tailed), double exponential (moderately heavy-tailed), and Cauchy (very heavy-tailed). Table 6.18 displays the proportion of times in 5,000 replications that the tests rejected H_0. The two populations were shifted by amounts 0, $.4\sigma$, and $.8\sigma$, where σ denotes the standard deviation of the distribution. (When the population distribution is Cauchy, σ denotes a scale parameter.)

TABLE 6.18
Empirical power of paired t (t) and signed-rank (T) tests with $\alpha = .05$

Distribution		Normal			Double Exponential			Cauchy			Uniform		
	Shift:	**0**	**.4σ**	**.8σ**	**0**	**.4σ**	**.8σ**	**0**	**.4σ**	**.8σ**	**0**	**.4σ**	**.8σ**
$n = 10$	t	.049	.330	.758	.047	.374	.781	.028	.197	.414	.051	.294	.746
	T	.050	.315	.741	.048	.412	.804	.049	.332	.623	.049	.277	.681
$n = 15$	t	.048	.424	.906	.049	.473	.898	.025	.210	.418	.051	.408	.914
	T	.047	.418	.893	.050	.532	.926	.050	.423	.750	.051	.383	.852
$n = 20$	t	.048	.546	.967	.044	.571	.955	.026	.214	.433	.049	.522	.971
	T	.049	.531	.962	.049	.652	.975	.049	.514	.849	.050	.479	.935

From Table 6.18, we can make the following observations. The level of the paired t test remains nearly equal to .05 for uniform and double exponential distributions, but is much less than .05 for the very heavy-tailed Cauchy distribution. The Wilcoxon signed-rank test's level is nearly .05 for all four distributions, as expected because the level of the Wilcoxon test requires only that the population distribution be symmetric. When the distribution is normal, the t test has only slightly greater power values than the Wilcoxon signed-rank test. When the population distribution is short-tailed and uniform, the paired t test has slightly greater power than the signed-rank test. Note also that the power values for the t test are slightly less than the t power values when the population distribution is normal. For the double exponential, the Wilcoxon test has slightly greater power than the t test. For the Cauchy distribution, the level of the t test deviates significantly from .05, and its power is much lower than that of the Wilcoxon test. From other studies, if the distribution of differences is grossly skewed, the nominal t probabilities may be misleading. The skewness has less of an effect on the level of the Wilcoxon test.

Even with this discussion, you might still be confused as to which statistical test or confidence interval to apply in a given situation. First, plot the data and attempt to determine whether the population distribution is very heavy-tailed or very skewed. In such cases, use a Wilcoxon rank-based test. When the plots are not definitive in their detection of nonnormality, perform both tests. If the results from the different tests yield different conclusions, carefully examine the data to identify any peculiarities to understand why the results differ. If the conclusions agree and there are no blatant violations of the required conditions, you should be very confident in your conclusions. This particular "hedging" strategy is appropriate not only for paired data but also for many situations in which there are several alternative analyses.

6.6 Choosing Sample Sizes for Inferences About $\mu_1 - \mu_2$

Sections 5.3 and 5.5 were devoted to sample-size calculations to obtain a confidence interval about μ with a fixed width and specified degree of confidence or to conduct a statistical test concerning μ with predefined levels for α and β. Similar calculations can be made for inferences about $\mu_1 - \mu_2$ with either independent samples or paired data. Determining the sample size for a $100(1-\alpha)\%$ confidence interval about $\mu_1 - \mu_2$ of width $2E$ based on independent samples is possible by solving the following expression for n:

$$z_{\alpha/2}\sigma\sqrt{\frac{1}{n}+\frac{1}{n}} = E$$

Note that, in this formula, σ is the common population standard deviation and we have assumed equal sample sizes.

Sample Sizes for a $100(1-\alpha)\%$ Confidence Interval for $\mu_1 - \mu_2$ of the Form $\bar{y}_1 - \bar{y}_2 \pm E$, Independent Samples

$$n = \frac{2z_{\alpha/2}^2\,\sigma^2}{E^2}$$

(*Note:* If σ is unknown, substitute an estimated value to get an approximate sample size.)

The sample sizes obtained using this formula are usually approximate because we have to substitute an estimated value of σ, the common population standard deviation. This estimate will probably be based on an educated guess from information on a previous study or on the range of population values.

Corresponding sample sizes for one- and two-sided tests of $\mu_1 - \mu_2$ based on specified values of α and β, where we desire a level α test having the probability of a Type II error $\beta(\mu_1 - \mu_2) \leq \beta$ whenever $|\mu_1 - \mu_2| \geq \Delta$, are shown here.

Sample Sizes for Testing $\mu_1 - \mu_2$, Independent Samples

$$\text{One-sided test:} \quad n = 2\sigma^2 \frac{(z_\alpha + z_\beta)^2}{\Delta^2}$$

$$\text{Two-sided test:} \quad n = 2\sigma^2 \frac{(z_{\alpha/2} + z_\beta)^2}{\Delta^2}$$

where $n_1 = n_2 = n$ and the probability of a Type II error is to be $\leq \beta$ when the true difference $|\mu_1 - \mu_2| \geq \Delta$. (*Note:* If σ is unknown, substitute an estimated value to obtain an approximate sample size.)

EXAMPLE 6.10

One of the crucial factors in the construction of large buildings is the amount of time it takes for poured concrete to reach a solid state, called the "set-up" time. Researchers are attempting to develop additives that will accelerate the set-up time without diminishing any of the strength properties of the concrete. A study is being designed to compare concrete with the most promising additive to concrete without the additive. The research hypothesis is that the concrete with the additive will have a smaller mean set-up time than the concrete without the additive. The researchers have decided to have the same number of test samples for the concrete with and without the additive. For an $\alpha = .05$ test, determine the appropriate number of test samples needed if we want the probability of a Type II error to be less than or equal to .10 whenever the concrete with the additive has a mean set-up time of 1.5 hours less than the concrete without the additive. From previous experiments, the standard deviation in set-up time is 2.4 hours.

Solution Let μ_1 be the mean set-up time for concrete without the additive and μ_2 be the mean set-up time for concrete with the additive. From the description of the problem, we have

- One-sided research hypothesis: $\mu_1 > \mu_2$
- $\sigma \approx 2.4$
- $\alpha = .05$
- $\beta \leq .10$ whenever $\mu_1 - \mu_2 \geq 1.5 = \Delta$
- $n_1 = n_2 = n$

From Table 1 in the Appendix, $z_\alpha = z_{.05} = 1.645$ and $z_\beta = z_{.10} = 1.28$. Substituting into the formula, we have

$$n \approx \frac{2\sigma^2(z_\alpha + z_\beta)^2}{\Delta^2} = \frac{2(2.4)^2(1.645 + 1.28)^2}{(1.5)^2} = 43.8, \text{ or } 44$$

Thus, we need 44 test samples of concrete with the additive and 44 test samples of concrete without the additive. ■

Sample-size calculations can also be performed when the desired sample sizes are unequal, $n_1 \neq n_2$. Let n_2 be some multiple m of n_1; that is, $n_2 = mn_1$. For example, we may want n_1 three times as large as n_2; hence, $n_2 = \frac{1}{3}n_1$. The displayed formulas can still be used, but we must substitute $(m + 1)/m$ for 2 and n_1 for n in the sample-size formulas. After solving for n_1, we have $n_2 = mn_1$.

EXAMPLE 6.11

Refer to Example 6.10. Because the set-up time for concrete without the additive has been thoroughly documented, the experimenters wanted more information about the concrete with the additive than about the concrete without the additive. In particular, the experimenters wanted three times more test samples of concrete with the additive than without the additive; that is, $n_2 = mn_1 = 3n_1$. All other specifications are as given in Example 6.10. Determine the appropriate values for n_1 and n_2.

Solution In the sample-size formula, we have $m = 3$. Thus, replace 2 with $\frac{m+1}{m} = \frac{4}{3}$. We then have

$$n_1 \approx \frac{\left(\frac{m+1}{m}\right)\sigma^2 (z_\alpha + z_\beta)^2}{\Delta^2} = \frac{\left(\frac{4}{3}\right)(2.4)^2(1.645 + 1.28)^2}{(1.5)^2} = 29.2, \text{ or } 30$$

Thus, we need $n_1 = 30$ test samples of concrete without the additive and $n_2 = mn_1 = (3)(30) = 90$ test samples with the additive. ■

Sample sizes for estimating μ_d and conducting a statistical test for μ_d based on paired data (differences) are found using the formulas of Chapter 5 for μ. The only change is that we are working with a single sample of differences rather than a single sample of y-values. For convenience, the appropriate formulas are shown here.

Sample Sizes for a 100(1 − α)% Confidence Interval for $\mu_1 - \mu_2$ of the Form $\bar{d} \pm E$, Paired Samples

$$n = \frac{z^2_{\alpha/2}\,\sigma_d^2}{E^2}$$

(*Note:* If σ_d is unknown, substitute an estimated value to obtain an approximate sample size.)

Sample Sizes for Testing $\mu_1 - \mu_2$, Paired Samples

One-sided test: $n = \dfrac{\sigma_d^2(z_\alpha + z_\beta)^2}{\Delta^2}$

Two-sided test: $n \cong \dfrac{\sigma_d^2(z_{\alpha/2} + z_\beta)^2}{\Delta^2}$

where the probability of a Type II error is β or less if the true difference $\mu_d \geq \Delta$. (*Note:* If σ_d is unknown, substitute an estimated value to obtain an approximate sample size.)

6.7 RESEARCH STUDY: Effects of an Oil Spill on Plant Growth

The oil company responsible for the oil spill described in the abstract at the beginning of this chapter implemented a plan to restore the marsh to prespill condition. To evaluate the effectiveness of the cleanup process, and in particular to study the

residual effects of the oil spill on the flora, researchers designed a study of plant growth 1 year after the burning. In an unpublished Texas A&M University dissertation, ***Newman (1998)*** describes the researchers' plan for evaluating the effect of the oil spill on *Distichlis spicata*, a flora of particular importance to the area of the spill. We will now describe a hypothetical set of steps that the researchers may have implemented in order to successfully design their research study.

Defining the Problem

The researchers needed to determine the important characteristics of the flora that may be affected by the spill. Some of the questions that needed to be answered prior to starting the study included the following:

1. What are the factors that determine the viability of the flora?
2. How did the oil spill affect these factors?
3. Are there data on the important flora factors prior to the spill?
4. How should the researchers measure the flora factors in the oil-spill region?
5. How many observations are necessary to confirm that the flora has undergone a change after the oil spill?
6. What type of experimental design or study is needed?
7. What statistical procedures are valid for making inferences about the change in flora parameters after the oil spill?
8. What types of information should be included in a final report to document the changes observed (if any) in the flora parameters?

Collecting the Data

The researchers determined that there was no specific information on the flora in this region prior to the oil spill. Since there was no relevant information on flora density in the spill region prior to the spill, it was necessary to evaluate the flora density in unaffected areas of the marsh to determine whether the plant density had changed after the oil spill. The researchers located several regions that had not been contaminated by the oil spill. They needed to determine how many tracts would be required in order for their study to yield viable conclusions. To determine how many tracts must be sampled, we have to determine how accurately the researchers want to estimate the difference in the mean flora densities in the spilled and unaffected regions. The researchers specified that they wanted the estimator of the difference in the two means to be within eight units of the true difference in the means. That is, the researchers wanted to estimate the difference in mean flora density with a 95% confidence interval having the form $y_{\text{Con}} - y_{\text{Spill}} \pm 8$. In previous studies on similar sites, the flora density ranged from 0 to 73 plants per tract. The number of tracts the researchers needed to sample in order to achieve their specifications would involve the following calculations.

We want a 95% confidence interval on $\mu_{\text{Con}} - \mu_{\text{Spill}}$ with $E = 8$ and $z_{\alpha/2} = z_{.025} = 1.96$. Our estimate of σ is $\hat{\sigma} = \text{range}/4 = (73 - 0)/4 = 18.25$. Substituting into the sample-size formula, we have

$$n = \frac{2(z_{\alpha/2})^2\hat{\sigma}^2}{E^2} = \frac{2(1.96)^2(18.25)^2}{(8)^2} = 39.98 \approx 40$$

Thus, a random sample of 40 tracts should give a 95% confidence interval for $\mu_{\text{Con}} - \mu_{\text{Spill}}$ with the desired tolerance of eight plants provided 18.25 is a reasonable estimate of σ.

The spill region and the unaffected regions were divided into tracts of nearly the same size. From the above calculations, it was decided that 40 tracts from both the spill and the unaffected areas would be used in the study. Forty tracts of exactly the same size were randomly selected in each of these locations, and the *Distichlis spicata* densities were recorded. The data consist of 40 measurements of flora density in the uncontaminated (control) sites and 40 density measurements in the contaminated (spill) sites. The data are given below in a stem-leaf plot. The researchers would next carefully examine the data from the fieldwork to determine if the measurements were recorded correctly. The data would then be transferred to computer files and prepared for analysis.

Summarizing Data

The next step in the study would be to summarize the data through plots and summary statistics. The data are displayed in Figure 6.8, with summary statistics given in Table 6.19. A boxplot of the data displayed in Figure 6.9 indicates that the control sites have a somewhat greater plant density than the oil-spill sites. From the summary statistics, we see that the average flora density in the control sites is $\bar{y}_{\text{Con}} = 38.48$ with a standard deviation of $s_{\text{Con}} = 16.37$. The sites within the spill region have an average density of $\bar{y}_{\text{Spill}} = 26.93$ with a standard deviation of $s_{\text{Spill}} = 9.88$. Thus, the control sites have a larger average flora density and a greater variability in flora density than do the sites within the spill region. Whether these observed differences in flora density reflect similar differences in all the sites and not just the ones included in the study will require a statistical analysis of the data.

FIGURE 6.8
Number of plants observed in tracts at oil-spill and control sites. The data are displayed in stem-and-leaf plots

Control Tracts				Oil-Spill Tracts		
Mean:	38.48	000	0		Mean:	26.93
Median:	41.50	7	0	59	Median:	26.00
St. Dev:	16.37	1	1	14	St. Dev:	9.88
n:	40	6	1	77799	*n*:	40
		4	2	2223444		
		9	2	555667779		
		0	3	11123444		
		55678	3	5788		
		000111222233	4	1		
		57	4			
		0112344	5	02		
		67789	5			

TABLE 6.19
Summary statistics for oil-spill data

Descriptive Statistics

Variable	Site Type	*N*	Mean	Median	Tr. Mean	St. Dev.
No. plants	Control	40	38.48	41.50	39.50	16.37
	Oil spill	40	26.93	26.00	26.69	9.88

Variable	Site Type	SE Mean	Minimum	Maximum	Q1	Q3
No. plants	Control	2.59	0.00	59.00	35.00	51.00
	Oil spill	1.56	5.00	52.00	22.00	33.75

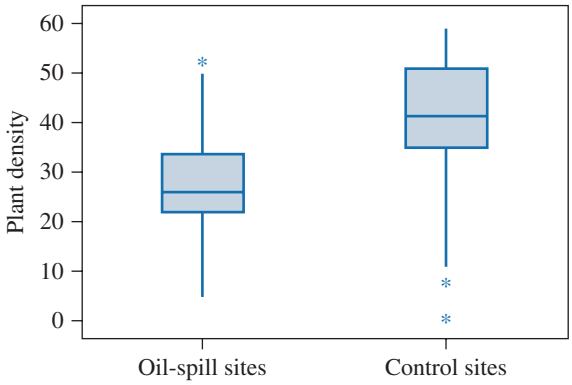


FIGURE 6.9
Number of plants observed in tracts at control sites (1) and oil-spill sites (2)

Analyzing Data

The researchers hypothesized that the oil-spill sites would have a lower plant density than the control sites. Thus, we will construct confidence intervals on the mean plant densities in the control plots, μ_{Con}, and in the oil-spill plots, μ_{Spill}, to assess their average plant density. Also, we can construct confidence intervals on the difference $\mu_{\text{Con}} - \mu_{\text{Spill}}$ and test the research hypothesis that μ_{Con} is greater than μ_{Spill}. From Figure 6.9, the data from the oil spill area appear to have a normal distribution, whereas the data from the control area appear to be skewed to the left. The normal probability plots are given in Figure 6.10 to further assess whether the population distributions are in fact normal in shape. We observe that the data from the spill tracts appear to follow a normal distribution but that the data from the control tracts do not, since their plotted points do not fall close to the straight line. Also, the variability in plant density is higher in the control sites than in the spill sites. Thus, the approximate t procedures will be the most appropriate inference procedures.

The sample data yielded the summary values shown in Table 6.20.

The research hypothesis is that the mean plant density for the control plots exceeds that for the oil-spill plots. Thus, our statistical test is set up as follows:

$$H_0: \mu_{\text{Con}} \le \mu_{\text{Spill}} \quad \text{versus} \quad H_a: \mu_{\text{Con}} > \mu_{\text{Spill}}$$

That is,

$$H_0: \quad \mu_{\text{Con}} - \mu_{\text{Spill}} \le 0$$
$$H_a: \quad \mu_{\text{Con}} - \mu_{\text{Spill}} > 0$$

$$\text{T.S.:} \quad t' = \frac{(\bar{y}_{\text{Con}} - \bar{y}_{\text{Spill}}) - D_0}{\sqrt{\dfrac{s^2_{\text{Con}}}{n_{\text{Con}}} + \dfrac{s^2_{\text{Spill}}}{n_{\text{Spill}}}}} = \frac{(38.48 - 26.93) - 0}{\sqrt{\dfrac{(16.37)^2}{40} + \dfrac{(9.88)^2}{40}}} = 3.82$$

In order to compute the rejection region and p-value, we need to compute the approximate df for t'.

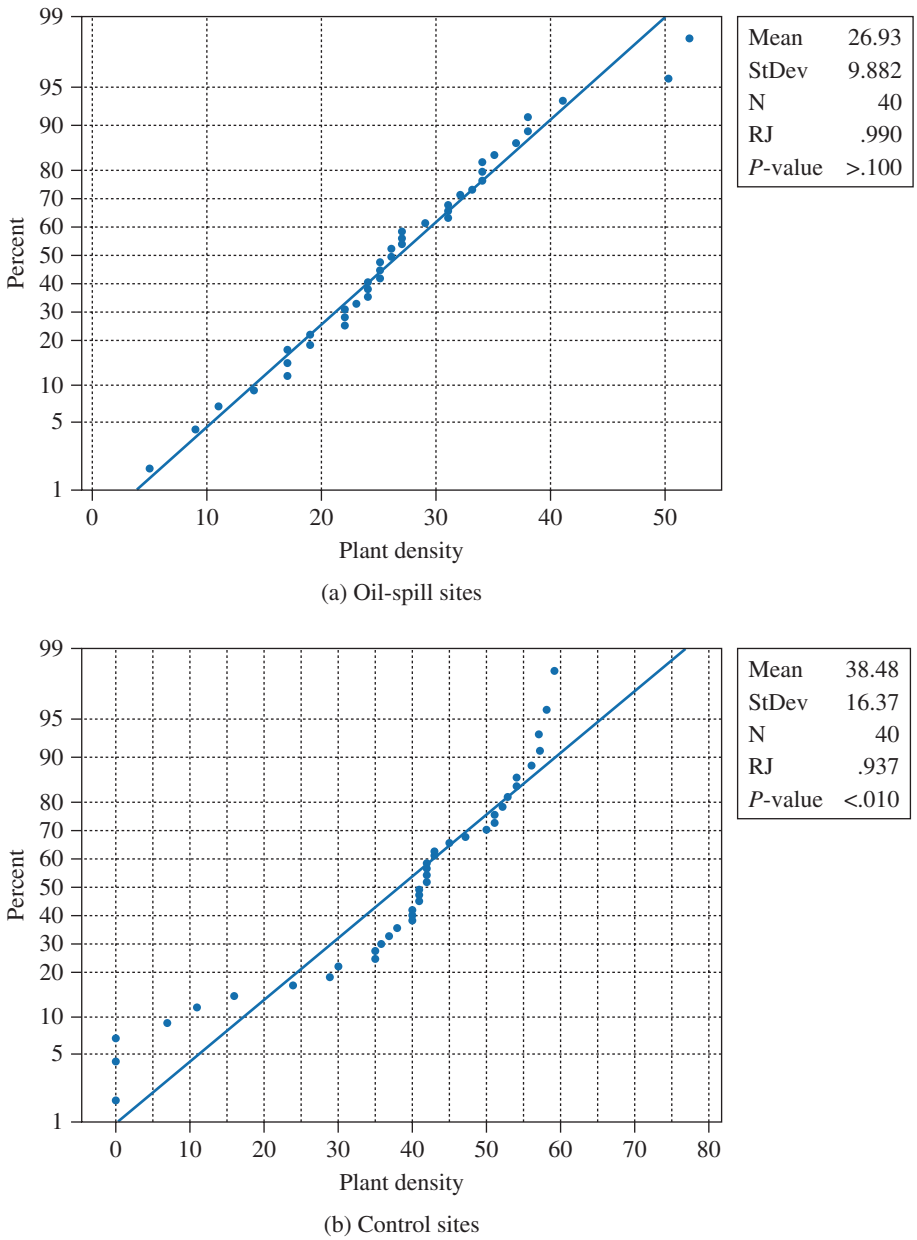
$$c = \frac{s^2_{\text{Con}}/n_{\text{Con}}}{\dfrac{s^2_{\text{Con}}}{n_{\text{Con}}} + \dfrac{s^2_{\text{Spill}}}{n_{\text{Spill}}}} = \frac{(16.37)^2/40}{(16.37)^2/40 + (9.88)^2/40} = .73$$

$$\text{df} = \frac{(n_{\text{Con}} - 1)(n_{\text{Spill}} - 1)}{(1 - c)^2(n_{\text{Con}} - 1) + c^2(n_{\text{Spill}} - 1)} = \frac{(39)(39)}{(1 - .73)^2(39) + (.73)^2(39)}$$
$$= 64.38, \text{ which is rounded to } 64$$

Since Table 2 in the Appendix does not have df = 64, we will use the R function $qt(1 - .05, 64) = 1.699$. In fact, the difference is very small when df becomes large: $t_{.05} = 1.671$ for df = 60, the value from Table 2.

R.R.: For $\alpha = .05$ and df = 64, reject H_0 if $t' > 1.699$.

FIGURE 6.10
Normal probability plots for the two types of sites



(a) Oil-spill sites

(b) Control sites

TABLE 6.20

Control Plots	Oil-Spill Plots
$n_{Con} = 40$	$n_{Spill} = 40$
$\bar{y}_{Con} = 38.48$	$\bar{y}_{Spill} = 26.93$
$s_{Con} = 16.37$	$s_{Spill} = 9.88$

Since $t' = 3.82$ is greater than 1.699, we reject H_0. We can bound the p-value using Table 2 in the Appendix with df = 60. With $t' = 3.82$, the level of significance is p-value $< .001$. Using R, p-value $= 1 - pt(3.82, 64) = .00015$. Thus, we can conclude that there is significant (p-value $< .00015$) evidence that μ_{Con} is greater than μ_{Spill}. Although we have determined that there is a statistically significant amount of evidence that the mean plant density at the control sites is greater than the mean plant density at the spill sites, the question remains whether these differences have *practical* significance. We can estimate the size of the difference in the means by placing a 95% confidence interval on $\mu_{\text{Con}} - \mu_{\text{Spill}}$.

The appropriate 95% confidence interval for $\mu_{\text{Con}} - \mu_{\text{Spill}}$ is computed by using the following formula with df = 64, the same as the value that was used for the R.R.

$$(\bar{y}_{\text{Con}} - \bar{y}_{\text{Spill}}) \pm t_{\alpha/2}\sqrt{\frac{s^2_{\text{Con}}}{n_{\text{Con}}} + \frac{s^2_{\text{Spill}}}{n_{\text{Spill}}}} =$$

$$(38.48 - 26.93) \pm 2.0\sqrt{\frac{(16.37)^2}{40} + \frac{(9.88)^2}{40}} = 11.55 \pm 6.05 = (5.5, 17.6)$$

Thus, we are 95% confident that the mean plant densities differ by an amount between 5.5 and 17.6. The plant scientists would then evaluate whether a difference in this range is of practical importance. This would then determine whether the sites in which the oil spill occurred have been returned to their prespill condition, at least in terms of this particular type of flora.

Reporting Conclusions

We would need to write a report summarizing our findings from the study. The following items should be included in the report:

1. Statement of objective for study
2. Description of study design and data collection procedures
3. Numerical and graphical summaries of data sets
   - table of means, medians, standard deviations, quartiles, range
   - boxplots
   - stem-and-leaf plots
4. Description of all inference methodologies:
   - approximate t tests of differences in means
   - approximate t-based confidence interval on population means
   - verification that all necessary conditions for using inference techniques were satisfied using boxplots, normal probability plots
5. Discussion of results and conclusions
6. Interpretation of findings relative to previous studies
7. Recommendations for future studies
8. Listing of data set

6.8 Summary and Key Formulas

In this chapter, we have considered inferences about $\mu_1 - \mu_2$. The first set of methods was based on independent random samples being selected from the populations of interest. We learned how to sample data to run a statistical test or to

construct a confidence interval for $\mu_1 - \mu_2$ using t methods. The Wilcoxon rank sum test, which does not require normality of the underlying populations, was presented as an alternative to the t test.

The second major set of procedures can be used to make comparisons between two populations when the sample measurements are paired. In this situation, we no longer have independent random samples, and, hence, the procedures of Sections 6.2 and 6.3 (t methods and the Wilcoxon rank sum test) are inappropriate. The test and estimation methods for paired data are based on the sample differences for the paired measurements or the ranks of the differences. The paired t test and corresponding confidence interval based on the difference measurements were introduced and found to be identical to the single-sample t methods of Chapter 5. The nonparametric alternative to the paired t test is the Wilcoxon signed-rank test.

The material presented in Chapters 5 and 6 lays the foundation of statistical inference (estimation and testing) for the remainder of the text. Review the material in this chapter periodically as new topics are introduced so that you retain the basic elements of statistical inference.

Key Formulas

1. $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$, independent samples; y_1 and y_2 approximately normal; $\sigma_1^2 = \sigma_2^2$

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad \text{and} \quad \text{df} = n_1 + n_2 - 2$$

2. t test for $\mu_1 - \mu_2$, independent samples; y_1 and y_2 approximately normal; $\sigma_1^2 = \sigma_2^2$

$$\text{T.S.: } t = \frac{\bar{y}_1 - \bar{y}_2 - D_0}{s_p \sqrt{1/n_1 + 1/n_2}} \quad \text{df} = n_1 + n_2 - 2$$

3. t' test for $\mu_1 - \mu_2$, unequal variances; independent samples; y_1 and y_2 approximately normal

$$\text{T.S.: } t' = \frac{\bar{y}_1 - \bar{y}_2 - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{df} = \frac{(n_1 - 1)(n_2 - 1)}{(1 - c)^2(n_1 - 1) + c^2(n_2 - 1)}$$

where

$$c = \frac{s_1^2/n_1}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

4. $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$, unequal variances; independent samples; y_1 and y_2 approximately normal

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where the t-percentile has

$$\text{df} = \frac{(n_1 - 1)(n_2 - 1)}{(1 - c)^2(n_1 - 1) + c^2(n_2 - 1)}$$

with

$$c = \frac{s_1^2/n_1}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$$

5. Wilcoxon rank sum test, independent samples

H_0: The two populations are identical.
$(n_1, n_2 \leq 10)$
T.S.: T, the sum of the ranks in sample 1
$(n_1, n_2 > 10)$

$$\text{T.S.:} \quad z = \frac{T - \mu_T}{\sigma_T}$$

where T denotes the sum of the ranks in sample 1

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \quad \text{and} \quad \sigma_T = \sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}$$

provided there are no tied ranks

6. Paired t test; differences approximately normal

$$\text{T.S.: } t = \frac{\bar{d} - D_0}{s_d/\sqrt{n}} \quad \text{df} = n - 1$$

where n is the number of differences

7. $100(1 - \alpha)\%$ confidence interval for μ_d, paired data; differences approximately normal

$$\bar{d} \pm t_{\alpha/2} s_d/\sqrt{n}$$

8. Wilcoxon signed-rank test, paired data

H_0: The distribution of differences is symmetrical about D_0.
T.S.: $(n \leq 50)$ T_- or T_+ or smaller of T_+ and T_- depending on the form of H_a
T.S.: $(n > 50)$

$$z = \frac{T - \mu_T}{\sigma_T}$$

where

$$\mu_T = \frac{n(n + 1)}{4} \quad \text{and} \quad \sigma_T = \sqrt{\frac{n(n + 1)(2n + 1)}{24}}$$

provided there are no tied ranks

9. Independent samples: sample sizes for estimating $\mu_1 - \mu_2$ with a $100(1 - \alpha)\%$ confidence interval, of the form $\bar{y}_1 - \bar{y}_2 \pm E$

$$n = \frac{2z_{\alpha/2}^2 \sigma^2}{E^2}$$

10. Independent samples: sample sizes for testing $\mu_1 - \mu_2$
    a. One-sided test:
$$n = \frac{2\sigma^2(z_\alpha + z_\beta)^2}{\Delta^2}$$
    b. Two-sided test:
$$n \cong \frac{2\sigma^2(z_{\alpha/2} + z_\beta)^2}{\Delta^2}$$
11. Paired samples: sample sizes for estimating $\mu_1 - \mu_2$ with $100(1 - \alpha)\%$ confidence interval, of the form $\bar{d} \pm E$
$$n = \frac{z_{\alpha/2}^2 \sigma_d^2}{E^2}$$
12. Paired samples: sample sizes for testing $\mu_1 - \mu_2$
    a. One-sided test:
$$n = \frac{\sigma_d^2(z_\alpha + z_\beta)^2}{\Delta^2}$$
    b. Two-sided test:
$$n \cong \frac{\sigma_d^2(z_{\alpha/2} + z_\beta)^2}{\Delta^2}$$

6.9 Exercises

6.1 Introduction

Env. **6.1** Refer to the oil-spill case study.
- a. What are the populations of interest?
- b. What are some factors other than flora density that may indicate that the oil spill has affected the marsh?
- c. Describe a method for randomly selecting the tracts where flora density measurements were to be taken.
- d. State several hypotheses that may be of interest to the researchers.

6.2 Inferences About $\mu_1 - \mu_2$: Independent Samples

Basic **6.2** For each of the situations, set up the rejection region:
- a. $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$ with $n_1 = 12, n_2 = 15$, and $\alpha = .05$
- b. $H_0: \mu_1 \leq \mu_2 + 3$ versus $H_a: \mu_1 > \mu_2 + 3$ with $n_1 = n_2 = 25$ and $\alpha = .01$
- c. $H_0: \mu_1 \geq \mu_2 - 9$ versus $H_a: \mu_1 < \mu_2 - 9$ with $n_1 = 13, n_2 = 15$, and $\alpha = .025$

Basic **6.3** Conduct a test of $H_0: \mu_1 \geq \mu_2 - 2.3$ versus $H_a: \mu_1 < \mu_2 - 2.3$ for the sample data summarized here. Use $\alpha = .01$ in reaching your conclusions.

	Population	
	1	2
Sample size	13	21
Sample mean	50.3	58.6
Sample standard deviation	7.23	6.98

Basic **6.4** Refer to Exercise 6.3.

a. What is the level of significance for your test?

b. Place a 99% confidence interval on $\mu_1 - \mu_2$.

Med. **6.5** In an effort to link cold environments with hypertension in humans, a preliminary experiment was conducted to investigate the effect of cold on hypertension in rats. Two random samples of 6 rats each were exposed to different environments. One sample of rats was held in a normal environment at 26°C. The other sample was held in a cold 5°C environment. Blood pressures and heart rates were measured for rats for both groups. The blood pressures for the 12 rats are shown in the accompanying table.

a. Do the data provide sufficient evidence that rats exposed to a 5°C environment have a higher mean blood pressure than rats exposed to a 26°C environment? Use $\alpha = .05$.

b. Evaluate the three conditions required for the test used in part (a).

c. Provide a 95% confidence interval on the difference in the two population means.

26°C		5°C	
Rat	**Blood Pressure**	**Rat**	**Blood Pressure**
1	152	7	384
2	157	8	369
3	179	9	354
4	182	10	375
5	176	11	366
6	149	12	423

Env. **6.6** The Department of Natural Resources (DNR) received a complaint from recreational fishermen that a community was releasing sewage into the river where they fished. These types of releases lower the level of dissolved oxygen in the river and hence cause damage to the fish residing in the river. An inspector from the DNR designs a study to investigate the fishermen's claim. Fifteen water samples are selected at locations on the river upstream from the community and fifteen samples are selected downstream from the community. The dissolved oxygen readings in parts per million (ppm) are given in the following table.

Upstream	5.2	4.8	5.1	5.0	4.9	4.8	5.0	4.7	4.7	5.0	4.6	5.2	5.0	4.9	4.7
Downstream	3.2	3.4	3.7	3.9	3.6	3.8	3.9	3.6	4.1	3.3	4.5	3.7	3.9	3.8	3.7

a. In order for the discharge to have an impact on fish health, there needs to be at least an .5 ppm reduction in the dissolved oxygen. Do the data provide sufficient evidence that there is a large enough reduction in the mean dissolved oxygen between the upstream and downstream water in the river to impact the health of the fish? Use $\alpha = .01$.

b. Do the required conditions to use the test in part (a) appear to be valid?

c. What is the level of significance of the test in part (a)?

d. Estimate the size of the difference in the mean dissolved oxygen readings for the two locations on the river using a 99% confidence interval.

Engin. **6.7** An industrial engineer conjectures that a major difference between successful and unsuccessful companies is the percentage of their manufactured products returned because of defectives. In a study to evaluate this conjecture, the engineer surveyed the quality control departments of 50 successful companies (identified by the annual profit statement) and 50 unsuccessful companies. The companies in the study all produced products of a similar nature and cost. The percentages of the total output returned by customers in the previous year are provided in following table.

a. Do the data provide sufficient evidence that successful businesses have a lower percentage of their products returned by customers? Use $\alpha = .05$.

Unsuccessful	11.35	9.19	10.30	8.59	4.98	6.82	6.03	11.15	9.38	8.32
Businesses	8.34	7.69	13.58	10.49	11.07	6.98	9.77	9.36	8.39	7.98
	6.56	6.85	8.06	7.71	11.04	11.69	9.40	10.00	5.45	9.67
	8.93	7.32	13.70	8.67	10.08	8.53	9.14	9.02	6.70	5.66
	8.26	7.07	12.23	11.93	4.76	13.81	11.41	6.44	9.50	8.99
Successful	10.24	6.16	5.06	10.64	6.77	10.13	4.59	1.38	8.81	1.97
Businesses	5.43	6.32	0.43	7.30	0.47	10.82	9.34	2.39	11.06	4.19
	5.09	8.20	10.51	1.94	9.82	6.69	0.91	6.17	0.17	7.47
	3.62	2.23	1.08	9.16	6.07	7.51	4.46	2.13	2.41	7.24
	4.06	7.70	8.32	6.33	3.83	4.96	9.05	6.41	0.27	8.48

b. Do the required conditions for applying your test in part (a) appear to be valid?

c. In order for the difference in percentage returns to have an economical impact, the difference must be at least 5%. Is there significant evidence that the percentage for successful businesses is at least 5% less that the percentage for unsuccessful businesses?

d. Estimate the difference in the percentages of returns for successful and unsuccessful businesses using a 95% confidence interval.

Soc. **6.8** The number of households currently receiving a daily newspaper has decreased over the last 10 years, and many people state they obtain information about current events through television news and the Internet. To test whether people who receive a daily newspaper have a greater knowledge of current events than people who don't, a sociologist gave a current events test to 25 randomly selected people who subscribe to a daily newspaper and to 30 randomly selected persons who do not receive a daily newspaper. The following stem-and-leaf graphs give the scores (maximum score is 70) for the two groups. Does it appear that people who receive a daily newspaper have a greater knowledge of current events? Be sure to evaluate all necessary conditions for your procedures to be valid.

```
Character Stem-and-Leaf Display

Stem-and-leaf of No Newspaper Deliver      Stem-and-leaf of Newspaper Subscribers
N=30                                       N=25
Leaf Unit = 1.0                            Leaf Unit = 1.0

        0 000
        0
        1 3
        1 59
        2 334                              2 2
        2 57                               2 99
        3 00234                            3 2
        3 5589                             3 66889
        4 00124                            4 000112333
        4 5                                4 55666

        5 0                                5 2
        5 55                               5 9
        6 2
```

Env. **6.9** The study of concentrations of atmospheric trace metals in isolated areas of the world has received considerable attention because of the concern that humans might somehow alter the climate of the earth by changing the amount and distribution of trace metals in the atmosphere. Consider a study at the South Pole, where, over a 2-month period, seventy air samples were obtained. In thirty-five of the samples, the amount of magnesium was determined. In the remaining thirty-five samples, the amount of europium was determined.

	Sample Size	Sample Mean	Sample Standard Deviation
Magnesium	35	1.0	2.21
Europium	35	17.0	12.65

a. What are the populations of interest in this study?
b. Is there significant evidence of a difference in the mean magnesium and Europium levels? Use $\alpha = .05$.
c. What is the level of significance of your test?
d. Estimate the mean levels of magnesium and Europium using a 95% confidence interval.

Env. **6.10** Refer to Exercise 6.9.
a. Based on the values of the sample mean and sample standard deviation for magnesium, provide a reason why the distribution of magnesium does not have a normal distribution.
b. Are the inferences given in Exercise 6.9 valid based on your answer in part (a)?

Env. **6.11** PCBs have been in use since 1929, mainly in the electrical industry, but it was not until the 1960s that they were found to be a major environmental contaminant. In the paper ***"The Ratio of DDE to PCB Concentrations in Great Lakes Herring Gull Eggs and Its Use in Interpreting Contaminants Data" [*Journal of Great Lakes Research (1998) *24(1):12–31]***, researchers report on the following study. Thirteen study sites from the five Great Lakes were selected. At each site, 9 to 13 herring gull eggs were collected randomly each year for several years. Following collection, the PCB content was determined. The mean PCB content at each site is reported in the following table for the years 1982 and 1996.

	Site												
Year	1	2	3	4	5	6	7	8	9	10	11	12	13
1982	61.48	64.47	45.50	59.70	58.81	75.86	71.57	38.06	30.51	39.70	29.78	66.89	63.93
1996	13.99	18.26	11.28	10.02	21.00	17.36	28.20	7.30	12.80	9.41	12.63	16.83	22.74

a. Legislation was passed in the 1970s restricting the production and use of PCBs. Thus, the active input of PCBs from current local sources has been severely curtailed. Do the data provide evidence that there has been a significant decrease in the mean PCB content of herring gull eggs?
b. Estimate the size of the decrease in mean PCB content from 1982 to 1996, using a 95% confidence interval.
c. Evaluate the conditions necessary to validly test the hypotheses and construct the confidence intervals using the collected data.
d. Does the independence condition appear to be violated?

6.12 Refer to Exercise 6.11. There appears to be a large variation in the mean PCB content across the 13 sites. How could we reduce the effect of variation in PCB content due to site differences on the evaluation of the difference in the PCB content means between the 2 years?

H.R. **6.13** A firm has a generous but rather complicated policy concerning end-of-year bonuses for its lower-level managerial personnel. The policy's key factor is a subjective judgment of "contribution to corporate goals." A personnel officer took samples of 24 female and 36 male managers to see whether there was any difference in bonuses, expressed as a percentage of yearly salary. The data are listed here:

Gender	Bonus Percentage								
F	9.2	7.7	11.9	6.2	9.0	8.4	6.9	7.6	7.4
	8.0	9.9	6.7	8.4	9.3	9.1	8.7	9.2	9.1
	8.4	9.6	7.7	9.0	9.0	8.4			
M	10.4	8.9	11.7	12.0	8.7	9.4	9.8	9.0	9.2
	9.7	9.1	8.8	7.9	9.9	10.0	10.1	9.0	11.4
	8.7	9.6	9.2	9.7	8.9	9.2	9.4	9.7	8.9
	9.3	10.4	11.9	9.0	12.0	9.6	9.2	9.9	9.0

a. What are the populations of interest in this study?
b. Is there significant evidence that the mean bonus percentage for males is more than five units larger than the mean bonus percentage for females? Use $\alpha = .05$.
c. What is the level of significance of your test?
d. Estimate the difference in the mean bonus percentages for males and females using a 95% confidence interval.

6.3 A Nonparametric Alternative: The Wilcoxon Rank Sum Test

Basic

6.14 Provide the rejection region for the Wilcoxon rank sum test for each of the following sets of hypotheses:

a. H_0: $\Delta = 0$ versus H_a: $\Delta \neq 0$ with $n_1 = 8$, $n_2 = 9$, and $\alpha = .10$
b. H_0: $\Delta = 0$ versus H_a: $\Delta < 0$ with $n_1 = 6$, $n_2 = 7$, and $\alpha = .05$
c. H_0: $\Delta = 0$ versus H_a: $\Delta > 0$ with $n_1 = 5$, $n_2 = 9$, and $\alpha = .025$

6.15 Random samples of size $n_1 = 8$ and $n_2 = 8$ were selected from populations A and B, respectively. The data are given in the following table.

Population A	4.3	4.6	4.7	5.1	5.3	5.3	5.8	5.4
Population B	3.5	3.8	3.7	3.9	4.4	4.7	5.2	4.4

a. Test for a difference in the medians of the two populations using an $\alpha = .05$ Wilcoxon rank sum test.
b. Place a 95% confidence interval on the difference in the medians of the two populations.

Basic

6.16 Refer to Exercise 6.15.

a. Test for a difference in the means in the two populations using an $\alpha = .05$ t-test.
b. Place a 95% confidence interval on the difference in the means of the two populations.
c. Compare the inferences obtained from the results from the Wilcoxon rank sum test and the t-test.
d. Which inferences appear to be more valid, inferences on the means or the medians?

Bus.

6.17 A cable TV company was interested in making its operation more efficient by cutting down on the distance between service calls while still maintaining at least the same level of service quality. A treatment group of 18 repairpersons was assigned to a dispatcher who monitored all the incoming requests for cable repairs and then provided a service strategy for that day's work orders. A control group of 18 repairpersons was to perform their work in a normal fashion—that is, by providing service in roughly a sequential order as requests for repairs were received. The average daily mileages for the 36 repairpersons are recorded here:

Treatment Group	62.2	79.3	83.2	82.2	84.1	89.3
	95.8	97.9	91.5	96.6	90.1	98.6
	85.2	87.9	86.7	99.7	101.1	88.6
Control Group	97.1	70.2	94.6	182.9	85.6	89.5
	109.5	101.7	99.7	193.2	105.3	92.9
	63.9	88.2	99.1	95.1	92.4	87.3

a. What are the populations of interest in this study?
b. Is there significant evidence that the treatment group had a smaller average daily mileage than the control group? Use $\alpha = .05$.
c. What is the level of significance of your test?
d. Estimate the difference in the average daily mileage for the treatment and control groups using a 95% confidence interval.
e. There are three possible procedures that could be applied to answer the questions in parts (b), (c), and (d). Which of these procedures appears to be the most valid?

Med. **6.18** The paper ***"Serum Beta-2-Microglobulin (SB2M) in Patients with Multiple Myeloma Treated with Alpha Interferon" [Journal of Medicine (1997) 28:311–318]*** reports on the influence of alpha interferon administration in the treatment of patients with multiple myeloma (MM). Twenty newly diagnosed patients with MM were entered into the study. The researchers randomly assigned the 20 patients to the two groups. Ten patients were treated with both intermittent melphalan and sumiferon (treatment group), whereas the remaining 10 patients were treated only with intermittent melphalan (control group). The SB2M levels were measured before and at days 3, 8, and 15 and months 1, 3, and 6 from the start of therapy. The measurement of SB2M was performed using a radioimmunoassay method. The measurements before treatment are given here.

Treatment Group	2.9	2.7	3.9	2.7	2.1	2.6	2.2	4.2	5.0	0.7
Control Group	3.5	2.5	3.8	8.1	3.6	2.2	5.0	2.9	2.3	2.9

a. Plot the sample data for both groups using boxplots or normal probability plots.
b. Based on your findings in part (a), which procedure appears more appropriate for comparing the distributions of SB2M?
c. Is there significant evidence that there is a difference in the distribution of SB2M for the two groups?
d. Discuss the implications of your findings in part (c) for the evaluation of the influence of alpha interferon.

6.19 The simulation study described in Section 6.3 evaluated the effect of heavy-tailed and skewed distributions on the level of significance and power of the t test and Wilcoxon rank sum test. Examine the results displayed in Table 6.13, and then answer the following questions.

a. What has a greater effect, if any, on the level of significance of the t test, skewness or heavy-tailness?
b. What has a greater effect, if any, on the level of significance of the Wilcoxon rank sum test, skewness or heavy-tailness?

6.20 Refer to Exercise 6.19.

a. What has a greater effect, if any, on the power of the t test, skewness or heavy tailedness?
b. What has a greater effect, if any, on the power of the Wilcoxon rank sum test, skewness or heavy tailedness?

6.21 Refer to Exercises 6.19 and 6.20.

a. For what type of population distributions would you recommend using the t test? Justify your answer.
b. For what type of population distributions would you recommend using the Wilcoxon rank sum test? Justify your answer.

6.4 Inferences About $\mu_1 - \mu_2$: Paired Data

Basic **6.22** Provide the rejection region for the paired t test for each of the following sets of hypotheses:

a. H_0: $\mu_d = 0$ versus H_a: $\mu_d \neq 0$ with $n = 19$, and $\alpha = .05$
b. H_0: $\mu_d \leq 0$ versus H_a: $\mu_d > 0$ with $n = 8$, and $\alpha = .025$
c. H_0: $\mu_d \geq 0$ versus H_a: $\mu_d < 0$ with $n = 14$, and $\alpha = .01$

Basic **6.23** A random sample of eight pairs of twins was randomly assigned to treatment A or treatment B. The data are given in the following table.

Twins	1	2	3	4	5	6	7	8
Treatment A	48.3	44.6	49.7	40.5	54.3	55.6	45.8	35.4
Treatment B	43.5	43.8	53.7	43.9	54.4	54.7	45.2	34.4

a. Is there significant evidence that the two treatments differ using an $\alpha = .05$ paired t test.
b. Is there significant evidence that the two treatments differ using an $\alpha = .05$ sign test.
c. Do your conclusions in parts (a) and (b) agree?
d. How do your inferences about the two treatments based on the paired t test and based on the sign test differ?

Basic **6.24** Refer to Exercise 6.23.

a. What is the level of significance of the paired t test?
b. What is the level of significance of the sign test?
c. Place a 95% confidence interval on the mean difference between the responses from the two treatments.
d. Which of the two procedures, the paired t test or the sign test, appears to be more valid in this study?

6.25 Refer to the data of Exercise 6.11. A potential criticism of analyzing these data as if they were two independent samples is that the measurements taken in 1996 were taken at the same sites as the measurements taken in 1982. Thus, there is the possibility that there will be a strong positive correlation between the pair of observations at each site.

a. Plot the pairs of observations in a scatterplot with the 1982 values on the horizontal axis and the 1996 values on the vertical axis. Does there appear to be a positive correlation between the pairs of measurements? Estimate the correlation between the pairs of observations?
b. Compute the correlation coefficient between the pairs of observations. Does this value confirm your observations from the scatterplot? Explain your answer.
c. Answer the questions posed in parts (a) and (b) of Exercise 6.11 using a paired data analysis. Are your conclusions different from the conclusions you reached treating the data as two independent samples?

Engin. **6.26** Researchers are studying two existing coatings used to prevent corrosion in pipes that transport natural gas. The study involves examining sections of pipe that had been in the ground at least 5 years. The effectiveness of the coating depends on the pH of the soil, so the researchers recorded the pH of the soil at all 20 sites at which the pipe was buried prior to measuring the amount of corrosion on the pipes. The pH readings are given here. Describe how the researchers could conduct the study to reduce the effect of the differences in the pH readings on the evaluation of the difference in the two coatings' corrosion protection.

	pH Readings at Twenty Research Sites									
Coating A	3.2	4.9	5.1	6.3	7.1	3.8	8.1	7.3	5.9	8.9
Coating B	3.7	8.2	7.4	5.8	8.8	3.4	4.7	5.3	6.8	7.2

Med. **6.27** Suppose you are a participant in a project to study the effectiveness of a new treatment for high cholesterol. The new treatment will be compared to a current treatment by recording the change in cholesterol readings over a 10-week treatment period. The effectiveness of the treatment may depend on each participant's age, body fat percentage, diet, and general health. The study will involve at most 30 participants because of cost considerations.

a. Describe how you would conduct the study using independent samples.
b. Describe how you would conduct the study using paired samples.
c. How would you decide which method, paired or independent samples, would be more efficient in evaluating the change in cholesterol readings?

Med. **6.28** The paper *"Effect of Long-Term Blood Pressure Control on Salt Sensitivity" [Journal of Medicine (1997) 28:147–156]* describes a study evaluating salt sensitivity (SENS) after a period of antihypertensive treatment. Ten hypertensive patients (diastolic blood pressure between 90 and 115 mmHg) were studied after at least 18 months on antihypertensive treatment. SENS readings, which were obtained before and after the patients were placed on an antihypertensive treatment, are given here.

Patient	1	2	3	4	5	6	7	8	9	10
Before treatment	22.86	7.74	15.49	9.97	1.44	9.39	11.40	1.86	−6.71	6.42
After treatment	6.11	−4.02	8.04	3.29	−0.77	6.99	10.19	2.09	11.40	10.70

a. Is there significant evidence that the mean SENS value decreased after the patient received antihypertensive treatment?
b. Estimate the size of the change in the mean SENS value.
c. Do the conditions required for using the t procedures appear to be valid for these data? Justify your answer.

Edu. **6.29** A study was designed to measure the effect of home environment on academic achievement of 12-year-old students. Because genetic differences may also contribute to academic achievement, the researcher wanted to control for this factor. Thirty sets of identical twins were identified who had been adopted prior to their first birthday, with one twin placed in a home in which academics were emphasized (Academic) and the other twin placed in a home in which academics were not emphasized (Nonacademic). The final grades (based on 100 points) for the 60 students are given here.

Set of Twins	Academic	Nonacademic	Set of Twins	Academic	Nonacademic
1	78	71	16	90	88
2	75	70	17	89	80
3	68	66	18	73	65
4	92	85	19	61	60
5	55	60	20	76	74
6	74	72	21	81	76
7	65	57	22	89	78
8	80	75	23	82	78
9	98	92	24	70	62
10	52	56	25	68	73
11	67	63	26	74	73
12	55	52	27	85	75
13	49	48	28	97	88
14	66	67	29	95	94
15	75	70	30	78	75

a. Is there a difference in the mean final grades between the students in an academically oriented home environment and those in a nonacademically oriented home environment. Use $\alpha = .05$.

b. Estimate the size of the difference in the mean final grades of the students in academic and nonacademic home environments using a 95% confidence interval.
c. Do the conditions for using the t procedures appear to be satisfied for these data?
d. Does it appear that using twins in this study to control for variation in final scores was effective as compared to taking a random sample of 30 students in both types of home environments? Justify your answer.

6.5 A Nonparametric Alternative: The Wilcoxon Signed-Rank Test

Basic **6.30** Provide the rejection region for the Wilcoxon signed-rank test for each of the following sets of hypotheses::

a. H_0: $M = 0$ versus H_a: $M \neq 0$ with $n = 19$, and $\alpha = .05$
b. H_0: $M \leq 0$ versus H_a: $M > 0$ with $n = 8$, and $\alpha = .025$
c. H_0: $M \geq 0$ versus H_a: $M < 0$ with $n = 14$, and $\alpha = .01$

Basic **6.31** A random sample of eight pairs of twins were randomly assigned to treatment A or treatment B. The data are given in the following table.

Twins	1	2	3	4	5	6	7	8
Treatment A	48.3	44.6	49.7	40.5	54.3	55.6	45.8	35.4
Treatment B	43.5	43.8	53.7	43.9	54.4	54.7	45.2	34.4

a. Is there significant evidence that the two treatments differ using an $\alpha = .05$ Wilcoxon signed-rank test.
b. Compare your conclusion with the conclusions obtained using the paired t test and sign test in Exercise 6.23.

Basic **6.32** Refer to Exercise 6.31.
a. What is the level of significance of the Wilcoxon signed-rank test?
b. Compare the levels of significance of the Wilcoxon signed-rank test, paired t test, and sign test for the data set in Exercise 6.31?
c. Place a 95% confidence interval on the mean difference between the responses from the two treatments.
d. Which of the three procedures, the Wilcoxon signed-rank test, paired t test or sign test, appears to be most valid test for this study?

6.33 Use the level and power values for the paired t test and Wilcoxon signed-rank test given in Table 6.18 to answer the following questions.
a. For small sample sizes, $n \leq 20$, does the actual level of the t test appear to deviate from the nominal level of $\alpha = .05$?
b. Which type of deviations from a normal distribution, skewness or heavy-tailedness, appears to have the greater affect on the t test?
c. For small sample sizes, $n \leq 20$, does the actual level of the Wilcoxon signed-rank test appear to deviate from the nominal level of $\alpha = .05$?
d. Which type of deviations from a normal distribution, skewness or heavy-tailedness, appears to have the greater effect on the Wilcoxon signed-rank test?

6.34 Use the level and power values for the paired t test and Wilcoxon signed-rank test given in Table 6.18 to answer the following questions:
a. Suppose a level .05 test is to be applied to a paired data set that has differences that are highly skewed to the right. Will the Wilcoxon signed-rank test's "actual" level or the paired t test's actual level be closer to .05? Justify your answer.
b. Suppose a boxplot of the differences in the pairs from a paired data set has many outliers, with an equal number above and below the median. If a level $\alpha = .05$ test is applied to the differences, will the Wilcoxon signed-rank test's "actual" level or the paired t test's actual level be closer to .05? Justify your answer.

Soc. **6.35** A study was conducted to determine whether automobile repair charges are higher for female customers than for male customers. Twenty auto repair shops were randomly selected from the telephone book. Two cars of the same age, brand, and engine problem were used in the study. For each repair shop, the two cars were randomly assigned to a man and woman participant and then taken to the shop for an estimate of repair cost. The repair costs (in dollars) are given here.

Repair Shop	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Female customers	871	684	795	838	1,033	917	1,047	723	1,179	707	817	846	975	868	1,323	791	1,157	932	1,089	770
Male customers	792	765	511	520	618	447	548	720	899	788	927	657	851	702	918	528	884	702	839	878

a. Which procedure, t or Wilcoxon, is more appropriate in this situation? Why?
b. Are repair costs generally higher for female customers than for male customers? Use $\alpha = .05$.

Bio. **6.36** The effect of Benzedrine on the heart rate of dogs (in beats per minute) was examined in an experiment on 14 dogs chosen for the study. Each dog was to serve as its own control, with half of the dogs assigned to receive Benzedrine during the first study period and the other half assigned to receive a placebo (saline solution). All dogs were examined to determine the heart rates after 2 hours on the medication. After 2 weeks in which no medication was given, the regimens for the dogs were switched for the second study period. The dogs previously on Benzedrine were given the placebo, and the others received Benzedrine. Again, heart rates were measured after 2 hours.

The following sample data are not arranged in the order in which they were taken but have been summarized by regimen. Use these data to test the research hypothesis that the distribution of heart rates for the dogs when receiving Benzedrine is shifted to the right of that for the same animals when on the placebo. Use a one-tailed Wilcoxon signed-rank test with $\alpha = .05$.

Dog	Placebo	Benzedrine	Dog	Placebo	Benzedrine
1	250	258	8	296	305
2	271	285	9	301	319
3	243	245	10	298	308
4	252	250	11	310	320
5	266	268	12	286	293
6	272	278	13	306	305
7	293	280	14	309	313

6.6 Choosing Sample Sizes for Inferences About $\mu_1 - \mu_2$

Med. **6.37** A study is being planned to evaluate the possible side effects of an anti-inflammatory drug. It is suspected that the drug may lead to an elevation in the blood pressure of users of the drug. A preliminary study of two groups of patients, one receiving the drug and the other receiving a placebo, provides the following information on the systolic blood pressure (in mm Hg) of the two groups:

Group	Mean	Standard Deviation
Placebo	129.9	18.5
Anti-inflammatory drug	135.5	18.7

Assume that both groups have systolic blood pressures that have a normal distribution with standard deviations relatively close to the values obtained in the pilot study. Suppose the study plan

provides for the same number of patients in the placebo group as in the treatment group. Determine the sample size necessary for an $\alpha = .05$ t test to have a power of .80 to detect an increase of 5 mm Hg in the blood pressure of the treatment group relative to that of the placebo group.

Med. **6.38** Refer to Exercise 6.37. Suppose that the agency sponsoring the study specifies that the group receiving the drug should have twice as many patients as the placebo group. Determine the sample sizes necessary for an $\alpha = .05$ t test to have a power of .80 to detect an increase of 5 mm Hg in the blood pressure of the treatment group relative to that of the placebo group.

Med. **6.39** Refer to Exercise 6.37. The researchers also need to obtain precise estimates of the mean difference in systolic blood pressures for people who use the anti-inflammatory drug versus those who do not.

a. Suppose the sample sizes are the same for both groups. What sample size is needed to obtain a 95% confidence interval for the mean difference in systolic blood pressure between the users and nonusers having a width of at most 5 mm Hg.

b. Suppose the user group will have twice as many patients as the placebo group. What sample size is needed to obtain a 95% confidence interval for the mean difference in systolic blood pressures between the users and nonusers having a width of at most 5 mm Hg.

Env. **6.40** An environmental impact study was performed in a small state to determine the effectiveness of scrubbers on the amount of pollution coming from the cooling towers of a chemical plant. The amounts of pollution (in ppm) detected from the cooling towers before and after the scrubbers were installed are given below for 23 cooling towers.

	Mean	Standard Deviation
Before scrubber	71	26
After scrubber	63	25
Difference = before − after	8	20

Suppose a larger study is planned for a state with a more extreme pollution problem.

a. How many chemical plant cooling towers need to be measured if we want a probability of .90 of detecting a mean reduction in pollution of 10 ppm due to installing the scrubbers using an $\alpha = .01$ test?

b. What assumptions did you make in part (a) in order to compute the sample size?

Env. **6.41** Refer to Exercise 6.40. The state regulators also need to obtain a precise estimate of the mean reduction in the pollution level after installing the scrubbers. What sample size is needed to obtain a 99% confidence interval having width of 8.5 ppm?

Supplementary Exercises

Med. **6.42** Long-distance runners have contended that moderate exposure to ozone increases lung capacity. To investigate this possibility, a researcher exposed 12 rats to ozone at the rate of two parts per million for a period of 30 days. The lung capacity of the rats was determined at the beginning of the study and again after the 30 days of ozone exposure. The lung capacities (in mL) are given here.

Rat	1	2	3	4	5	6	7	8	9	10	11	12
Before exposure	8.7	7.9	8.3	8.4	9.2	9.1	8.2	8.1	8.9	8.2	8.9	7.5
After exposure	9.4	9.8	9.9	10.3	8.9	8.8	9.8	8.2	9.4	9.9	12.2	9.3

a. Is there sufficient evidence to support the conjecture that ozone exposure increases lung capacity? Use $\alpha = .05$. Report the p-value of your test.
b. Estimate the size of the increase in lung capacity after exposure to ozone using a 95% confidence interval.
c. After completion of the study, the researcher claimed that ozone causes increased lung capacity. Is this statement supported by this experiment?

Env. **6.43** In an environmental impact study for a new airport, the noise levels of various jets were measured just seconds after their wheels left the ground. The jets were either wide-bodied or narrow-bodied. The noise levels in decibels (dB) are recorded here for 15 wide-bodied jets and 12 narrow-bodied jets.

Wide-Bodied Jets	109.5	107.3	105.0	117.3	105.4	113.7	121.7	109.2	108.1	106.4	104.6	110.5	110.9	111.0	112.4
Narrow-Bodied Jets	131.4	126.8	114.1	126.9	108.2	122.0	106.9	116.3	115.5	111.6	124.5	116.2			

a. Do the two types of jets have different mean noise levels? Report the level of significance of the test.
b. Estimate the size of the difference in mean noise levels between the two types of jets using a 95% confidence interval.
c. How would you select the jets for inclusion in this study?

Ag. **6.44** An entomologist is investigating which of two fumigants, F_1 or F_2, is more effective in controlling parasities in tobacco plants. To compare the fumigants, nine fields of differing soil characteristics, drainage, and amount of wind shield were planted with tobacco. Each field was then divided into two plots of equal area. Fumigant F_1 was randomly assigned to one plot in each field and F_2 to the other plot. Fifty plants were randomly selected from each field, 25 from each plot, and the numbers of parasites were counted. The data are in the following table.

Field	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Fumigant F_1	77	40	11	31	28	50	53	26	33
Fumigant F_2	76	38	10	29	27	48	51	24	32

a. What are the populations of interest?
b. Do the data provide sufficient evidence to indicate a difference in the mean levels of parasites for the two fumigants? Use $\alpha = .10$. Report the p-value for the experimental data.
c. Estimate the size of the difference in the mean numbers of parasites between the two fumigants using a 90% confidence interval.

6.45 Refer to Exercise 6.44. An alternative design of the experiment would involve randomly assigning fumigant F_1 to nine of the plots and F_2 to the other nine plots, ignoring which fields the plots were from. What are some of the problems that may occur in using the alternative design?

Env. **6.46** Following the March 24, 1989, grounding of the tanker *Exxon Valdez* in Alaska, approximately 35,500 tons of crude oil were released into Prince William Sound. The paper ***"The Deep Benthos of Prince William Sound, Alaska, 16 Months After the* Exxon Valdez *Oil Spill" (Feder and Blanchard, 1998)*** reports on an evaluation of deep benthic infauna after the spill. Thirteen sites were selected for study. Seven of the sites were within the oil trajectory, and six were outside the oil trajectory. Collection of environmental and biological data at two depths,

40 m and 100 m, occurred in the period July 1–23, 1990. One of the variables measured was population abundance (individuals per square meter). The values are given in the following table.

	Within Oil Trajectory							Outside Oil Trajectory					
Site	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**
Depth 40 m	5,124	2,904	3,600	2,880	2,578	4,146	1,048	1,336	394	7,370	6,762	744	1,874
Depth 100 m	3,228	2,032	3,256	3,816	2,438	4,897	1,346	1,676	2,008	2,224	1,234	1,598	2,182

a. After combining the data from the two depths, does there appear to be a difference in population mean abundances between the sites within and outside the oil trajectory? Use $\alpha = .05$.
b. Estimate the size of the difference in the mean population abundances at the two types of sites using a 95% confidence interval.
c. What are the required conditions for the techniques used in parts (a) and (b)?
d. Check to see whether the required conditions are satisfied.

6.47 Refer to Exercise 6.46. Answer the following questions using the combined data for both depths.

a. Use the Wilcoxon rank sum test to assess whether there is a difference in population abundances between the sites within and outside the oil trajectory. Use $\alpha = .05$.
b. What are the required conditions for the techniques used in part (a)?
c. Are the required conditions satisfied?
d. Discuss any differences in the conclusions obtained using the t procedures and the Wilcoxon rank sum test.

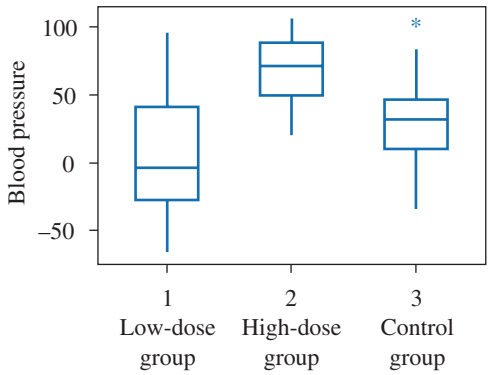
6.48 Refer to Exercise 6.46. The researchers also examined the effect of depth on population abundance.

a. Plot the four data sets using side-by-side boxplots to demonstrate the effect of depth on population abundance.
b. Separately for each depth, evaluate differences between the sites within and outside the oil trajectory. Use $\alpha = .05$.
c. Are your conclusions at 40 m consistent with your conclusions at 100 m?

6.49 Refer to Exercises 6.46–6.48.

a. Discuss the veracity of the following statement: "The oil spill did not adversely affect the population abundance; in fact, it appears to have increased the population abundance."
b. A possible criticism of the study is that the six sites outside the oil trajectory were not comparable in many aspects to the seven sites within the oil trajectory. Suppose that the researchers had data on population abundance at the seven within-trajectory sites prior to the oil spill. What type of analysis could be used on these data to evaluate the effect of the oil spill on population abundance? What are some advantages to using these data rather than the data in Exercise 6.46?
c. What are some possible problems with using the before and after oil spill data in assessing the effect of the spill on population abundance?

Bio. **6.50** A study was conducted to evaluate the effectiveness of an antihypertensive product. Three groups of 20 rats each were randomly selected from a strain of hypertensive rats. The 20 rats in the first group were treated with a low dose of an antihypertensive product, the second group with a higher dose of the same product, and the third group with an inert control. The amounts of decrease in systolic blood pressure 30 minutes after the rats receive an injection are given in the following table. Note that negative values represent increases in blood pressure.



Row	Low Dose	High Dose	Control
1	−45.1	54.2	18.2
2	−59.8	89.1	17.2
3	58.1	89.6	34.8
4	−23.7	98.8	3.2
5	64.9	107.3	42.9
6	12.1	65.1	−27.2
7	10.5	75.6	42.6
8	42.5	52.0	10.0
9	48.5	50.2	102.3
10	−1.7	80.9	61.0
11	−65.4	92.6	−33.1
12	−17.5	55.3	55.1
13	22.1	103.2	84.6
14	−15.4	45.4	40.3
15	96.5	70.9	30.5
16	−27.7	29.7	18.5
17	−16.7	40.3	29.3
18	39.5	73.3	−19.7
19	−4.2	21.0	37.2
20	−41.3	73.2	48.8

---

a. Compare the mean drops in blood pressure for the high-dose group and the control group. Use $\alpha = .05$ and report the level of significance.

b. Estimate the size of the difference in the mean drops for the high-dose and control groups using a 95% confidence interval.

c. Do the conditions required for the statistical techniques used in parts (a) and (b) appear to be satisfied? Justify your answer.

6.51 Refer to Exercise 6.50.

a. Compare the mean drops in blood pressure for the low-dose group and the control group. Use $\alpha = .05$ and report the level of significance.

b. Estimate the size of the difference in the mean drops for the low-dose and control groups using a 95% confidence interval.

c. Do the conditions required for the statistical techniques used in parts (a) and (b) appear to be satisfied? Justify your answer.

6.52 Refer to Exercise 6.50.

a. Compare the mean drops in blood pressure for the low-dose group and the high-dose group. Use $\alpha = .05$ and report the level of significance.

b. Estimate the size of the difference in the mean drops for the low-dose and high-dose groups using a 95% confidence interval.

c. Do the conditions required for the statistical techniques used in parts (a) and (b) appear to be satisfied? Justify your answer.

Med. **6.53** Refer to Exercise 6.50.

a. Describe the populations to which the inferences provided in Exercises 6.50–6.52 are relevant.

b. A much larger study is to be designed to further examine the effectiveness of the high-dose level of the drug. How many rats would be needed in the new study to be 90% confident that an $\alpha = .05$ test would detect a reduction of 10 mm Hg by the high-dose level relative to the mean blood pressure readings of the control group? *Hint:* Assume that the decreases in blood pressure for the high-dose and control groups have normal distributions with standard deviations of 30 mm Hg.

c. The company producing the drug wants a precise estimate of the mean reduction in the systolic blood pressure after injection with a high dose of the drug. What sample size is needed to obtain a 99% confidence interval having width of 5 mm Hg?

Med. **6.54** To assess whether degreed nurses received a more comprehensive training than registered nurses, a study was designed to compare the two groups. The state nursing licensing board randomly selected 50 nurses from each group for evaluation. They were given the state licensing board examination, and their scores are given in the following table.

Degreed	429	408	418	402	424	369	372	406	391	404
Nurses	408	417	422	408	365	412	379	423	412	420
	382	394	399	403	373	434	406	428	398	418
	383	395	408	402	416	424	439	382	371	386
	382	404	381	430	394	410	382	410	394	404
Registered	364	330	368	342	327	310	347	361	364	358
Nurses	362	356	333	347	356	306	375	345	420	332
	354	390	382	342	348	389	354	338	328	339
	320	382	295	341	387	284	383	311	387	397
	363	365	309	327	321	352	416	380	341	330

a. Can the licensing board conclude that the mean score of nurses who receive a BS in nursing is higher than the mean score of registered nurses? Use $\alpha = .05$.

b. Report the p-value for your test.

c. Estimate the size of the difference in the mean scores of the two groups of nurses using a 95% confidence interval.

d. The mean test scores are considered to have a meaningful difference only if they differ by more than 40 points. Is the observed difference in the mean scores a meaningful one?

Pol. Sci. **6.55** All persons running for public office must report the amounts of money spent during their campaigns. Political scientists have contended that female candidates generally find it difficult to raise money and therefore spend less in their campaigns than do male candidates. Suppose the accompanying data represent the campaign expenditures of a randomly selected group of 20 male and 20 female candidates for the state legislature. Do the data support the claim that female candidates generally spend less in their campaigns for public office than do male candidates?

Campaign Expenditures (in thousands of dollars)

Candidate	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Female	169	206	257	294	252	283	240	207	230	183	298	269	256	277	300	126	318	184	252	305
Male	289	334	278	268	336	438	388	388	394	394	425	386	356	342	305	365	355	312	209	458

a. Estimate the size of the difference in the mean campaign expenditures between female and male candidates using a 95% confidence interval.

b. Is there a significant difference at the .05 level in the mean campaign expenditures between female and male candidates?

c. Is there a practical difference in the mean campaign expenditures between female and male candidates?

d. Are the conditions necessary to analyze the data using the t test to satisfied?

Pol. Sci. **6.56** Refer to Exercise 6.55.

a. To what populations are the conclusions obtained in Exercise 6.55 relevant?

b. A more precise estimate of the mean expenditure for female candidates is requested. How many female candidates would need to be included in the new study to estimate the mean expenditure using a 95% confidence interval having a width of at most $10?

Env. **6.57** After strip-mining for coal, the state land office requires the mining company to restore the land to its condition prior to mining. One of many factors that is considered is the pH of the soil, which is an important factor in determining what types of plants will survive in a given location. The area to be mined was divided into grids before the mining took place. Fifteen grids were randomly selected, and the soil pH was measured before mining. When the mining was completed, the land was restored, and another set of pH readings was taken on the same 15 grids; see the accompanying table.

Location	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	10.02	10.16	9.96	10.01	9.87	10.05	10.07	10.08	10.05	10.04	10.09	10.09	9.92	10.05	10.13
After	10.21	10.16	10.11	10.10	10.07	10.13	10.08	10.30	10.17	10.10	10.06	10.37	10.24	10.19	10.13

a. What is the level of significance of the test for a change in mean pH after reclamation of the land?

b. What is the research hypothesis that the land office was testing?

c. Estimate the change in mean soil pH after strip-mining using a 99% confidence interval.

d. The land office assessed a fine on the mining company because the t test indicated a significant difference in mean pH after the reclamation of the land. Is the assessment of the fine supported by the data? Justify your answer using the results from parts (a) and (c).

6.58 Refer to Exercise 6.57. Based on the land office's decision in the test of hypotheses, could it have made (select one of the following)

a. A Type I error?

b. A Type II error?

c. Both a Type I and a Type II error?

d. Neither a Type I nor a Type II error?

Med. **6.59** Company officials are concerned about the length of time a particular drug retains its potency. A random sample (sample 1) of 10 bottles of the product is drawn from current production and analyzed for potency. A second sample (sample 2) is obtained, stored for 1 year, and then analyzed. The readings obtained are as follows:

Sample 1	10.2	10.5	10.3	10.8	9.8	10.6	10.7	10.2	10.0	10.6
Sample 2	9.8	9.6	10.1	10.2	10.1	9.7	9.5	9.6	9.8	9.9

a. What is the research hypothesis?

b. Compute the values of the t and t' statistics? Why are they equal for this data set?

c. What are the p-values for the t and t' statistics? Why are they different?

d. Are the conclusions concerning the research hypothesis the same for the two tests if we use $\alpha = .05$?

e. Which test, t or t', is more appropriate for this data set?

Engin. **6.60** An industrial concern has experimented with several different mixtures of the four components—magnesium, sodium nitrate, strontium nitrate, and a binder—that comprise a rocket propellant. The company has found that two mixtures in particular give higher flare-illumination values than the others. Mixture 1 consists of a blend composed of the proportions .40, .10, .42, and .08, respectively, for the four components of the mixture; mixture 2 consists of a blend using the proportions .60, .25, .10, and .05. Twenty different blends (10 of each mixture) are prepared and tested to obtain the flare-illumination values. These data appear here (in units of 1,000 candles).

Mixture 1	185	192	201	215	170	190	175	172	198	202
Mixture 2	221	210	215	202	204	196	225	230	214	217

a. Plot the sample data. Which test(s) could be used to compare the mean illumination values for the two mixtures?

b. Give the level of significance of the test and interpret your findings.

6.61 Refer to Exercise 6.60. Instead of conducting a statistical test, use the sample data to answer the question, What is the difference in mean flare illuminations for the two mixtures?

6.62 Refer to Exercise 6.60. Suppose we wish to test the research hypothesis that $\mu_1 < \mu_2$ for the two mixtures. Assume that the population distributions are normally distributed with a common $\sigma = 12$. Determine the sample size required to obtain a test having $\alpha = .05$ and $\beta(\mu_d) < .10$ when $\mu_2 - \mu_1 \geq 15$.

Med. **6.63** Refer to the epilepsy study data in Table 3.19. Use the data for the number of seizures after 8 weeks for the placebo patients and for the patients treated with the drug progabide to answer the following questions.

a. Do the data support the conjecture that progabide reduces the mean number of seizures for epileptics? Use both a t test and the Wilcoxon test with $\alpha = .05$.

b. Which test appears to be more appropriate for this study? Why?

c. Estimate the size of the difference in the mean numbers of seizures between the two groups.

Bus. **6.64** Many people purchase sport utility vehicles (SUVs) because they think they are sturdier and hence safer than regular cars. However, preliminary data have indicated that the costs for repairs of SUVs are higher than for midsize cars when both vehicles are in an accident. A random sample of 8 new SUVs and 8 midsize cars is tested for front-impact resistance. The amounts of damage (in hundreds of dollars) to the vehicles when crashed at 20 mph head on into a stationary barrier are recorded in the following table.

Car	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
SUV	14.23	12.47	14.00	13.17	27.48	12.42	32.59	12.98
Midsize	11.97	11.42	13.27	9.87	10.12	10.36	12.65	25.23

a. Plot the data to determine whether the conditions required for the t procedures are valid.

b. Do the data support the conjecture that the mean damage is greater for SUVs than for midsize vehicles? Use $\alpha = .05$ with both the t test and the Wilcoxon test.

c. Which test appears to be the more appropriate procedure for this data set?

d. Do you reach the same conclusions from both procedures? Why or why not?

6.65 Refer to Exercise 6.64. The small number of vehicles in the study has led to criticism of the results. A new study is to be conducted with a larger sample size. Assume that both populations of damages are normally distributed with a common $\sigma = \$700$.

a. Determine the sample size that allows us to be 95% confident that the estimate of the difference in mean repair costs is within \$500 of the true difference.

b. For the research hypothesis H_a: $\mu_{SUV} > \mu_{MID}$, determine the sample size required to obtain a test having $\alpha = .05$ and $\beta(\mu_d) < .05$ when $\mu_{SUV} - \mu_{MID} \geq \$500$.

Law **6.66** The following memorandum opinion on statistical significance was issued by the judge in a trial involving many scientific issues. The opinion has been stripped of some legal jargon and has been taken out of context. Still, it can give us an understanding of how others deal with the problem of ascertaining the meaning of statistical significance. Read this memorandum and comment on the issues raised regarding statistical significance.

Memorandum Opinion

This matter is before the Court upon two evidentiary issues that were raised in anticipation of trial. First, it is essential to determine the appropriate level of statistical significance for the admission of scientific evidence.

With respect to statistical significance, no statistical evidence will be admitted during the course of the trial unless it meets a confidence level of 95%.

Every relevant study before the Court has employed a confidence level of at least 95%. In addition, plaintiffs concede that social scientists routinely utilize a 95% confidence level. Finally, all legal authorities agree that statistical evidence is admissible only if it meets the 95% confidence level required by statisticians. Therefore, because plaintiffs advance no reasonable basis to alter the accepted approach of mathematicians to the test of statistical significance, no statistical evidence will be admitted at trial unless it satisfies the 95% confidence level.

Env. **6.67** **Defining the Problem (1).** Lead is an environmental pollutant especially worthy of attention because of its damaging effects on the neurological and intellectual development of children. Morton et al. (1982) collected data on lead absorption by children whose parents worked at a factory in Oklahoma where lead was used in the manufacture of batteries. The concern was that children might be exposed to lead inadvertently brought home on the bodies or clothing of their parents. Levels of lead (in micrograms per deciliter) were measured in blood samples taken from 33 children who might have been exposed in this way. They constitute the exposed group.

Collecting the Data (2). The researchers formed a control group by making matched pairs. For each of the 33 children in the exposed group they selected a matching child of the same age, living in the same neighborhood, and with parents employed at a place where lead is not used.

The data set LEADKIDS contains three variables, each with 33 cases. All involve measurements of lead in micrograms per deciliter of blood.

```
c1   Exposed   Lead (μg/dl of whole blood) for children of workers
                in the battery factory
c2   Control   Lead (μg/dl of whole blood) for matched controls
c3   Diff      The differences: 'Exposed' - 'Control'.
```

These data are listed next.

LEADKIDS

Exposed	Control	Diff
38	16	22
23	18	5
41	18	23
18	24	-6
37	19	18
36	11	25
23	10	13
62	15	47
31	16	15
34	18	16
24	18	6
14	13	1
21	19	2
17	10	7
16	16	0
20	16	4
15	24	-9
10	13	-3
45	9	36
39	14	25
22	21	1
35	19	16
49	7	42
48	18	30
44	19	25
35	12	23
43	11	32
39	22	17
34	25	9
13	16	-3
73	13	60
25	11	14
27	13	14

LEADEXP

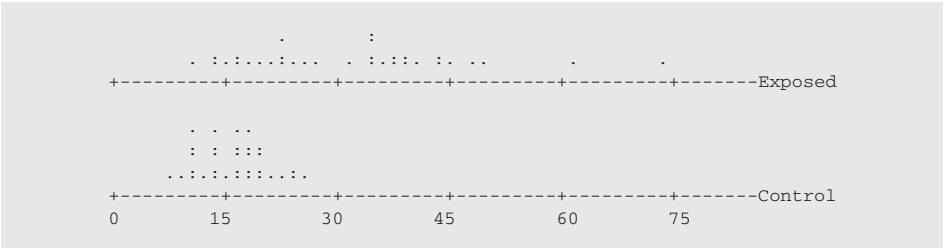
Lead	JobExp	JobHyg
14	3	1
13	3	1
25	3	1
39	2	1
41	3	2
18	3	2
49	3	2
29	2	2
16	1	2
38	3	3
23	3	3
37	3	3
62	3	3
24	3	3
45	3	3
39	3	3
48	3	3
44	3	3
35	3	3
43	3	3
34	3	3
73	3	3
31	2	3
34	2	3
20	2	3
22	2	3
35	2	3
36	1	3
23	1	3
21	1	3
17	1	3
27	1	3
15	1	3
10	1	3

This is necessarily an observational study rather than a controlled experiment. There is no way that the researchers could have assigned children at random to parents in or out of lead-related occupations. Furthermore, the exposed subjects were all chosen from the small group of children whose parents worked at one particular plant. They were not chosen from the larger population of children everywhere who might be exposed to lead as a result of their parents' working conditions.

If lead levels are unusually high in the exposed group, it might be argued that the lead in their blood came from some source other than their parents' place of work: from lead solder in water pipes at home, from lead-paint dust at school, from air pollution, and so on. For this reason, a properly chosen control group of children is crucial to the credibility of the study.

In principle, the children in the control group should be subject to all of the same possible lead contaminants as those in the exposed group except for lead brought home from work by parents. In practice, the designers of this study chose to use two criteria in forming pairs: neighborhood and age. Neighborhood seems a reasonable choice because general environmental conditions, types of housing, and so on could vary greatly for children living in different neighborhoods. Controlling for age seems reasonable because lead poisoning is largely cumulative, so levels of lead might be higher in older children. Thus, for each child in the exposed group, researchers sought a paired child of the same age and living in the same neighborhood.

Summarizing the Data (3). We begin by looking at dot plots of the data for the exposed and control groups:

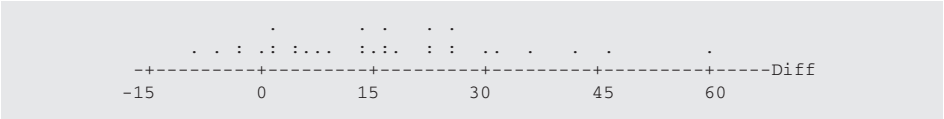


We can see that over half of the children in the exposed group have more lead in their blood than do any of the children in the control group. This graphical comparison is not the most effective one we could make because it ignores the pairing of exposed and control children. Even so, it presents clear evidence that, on average, the exposed children have more lead in their blood than do the control children.

Notice that the lead levels of the exposed group are much more diverse than those of the control group. This suggests that some children in the exposed group are getting a lot more lead, presumably from their working parents, than are others in this group. Perhaps some parents at the battery factory do not work in areas where they come into direct contact with lead. Perhaps some parents wear protective clothing that is left at work, or they shower before they leave work. For this study, information on the exposure and hygiene of parents was collected by the investigators. Such factors were found to contribute to the diversity of the lead levels observed among the exposed children.

Some toxicologists believe that *any* amount of lead may be detrimental to children, but all agree that the highest levels among the exposed children in our study are dangerously high. Specifically, it is generally agreed that children with lead levels above 40 micrograms per deciliter need medical treatment. Children above 60 on this scale should be immediately *hospitalized* for treatment ***(Miller and Keane, 1957)***. A quick glance at the dot plot shows that we are looking at some serious cases of lead poisoning in the exposed group.

By plotting differences, we get an even sharper picture. For each matched pair of children the variable `Diff` shows how much more lead the exposed child has than his or her control neighbor of the same age.



If we consider a hypothetical population of pairs of children, the difference measures the increased lead levels that may result from exposure via a parent working at the battery factory.

If parents who work at the battery factory were not bringing lead home with them, we would expect about half of these values to be positive and half to be negative. The lead values in the blood would vary but in such a way that the exposed child would have only a 50–50 chance of having the higher value. Thus, we would expect the dot plot to be centered near 0.

In contrast, look at the dot plot of the actual data. Almost every child in the exposed group has a higher lead value than does the corresponding control child. As a result, most of the differences are positive. The average of the differences is the *balance point* of the dot plot, located somewhat above 15. (In some respects, we can read the dot plot quite precisely. In 1 pair out of 33, both children have the same value, to the nearest whole number as reported. In only 4 pairs does the control child have the higher level of lead.)

The dot plot of the differences displays strong evidence that the children in the exposed group have more lead than their control counterparts. It will be necessary to perform some formal statistical tests to check whether this effect is statistically significant, but we already suspect from this striking graph what the conclusion must be.

We have looked directly at the *pairs* of children around which the study was built. It may take a bit more thought to deal with differences than to look at the separate variables exposed and control as we did previously. But looking at pairs is best. If the effect had turned out to be weaker and if we had not thought to look at pairs, then we might have missed seeing the effect.

a. Obtain the mean, median, and standard deviation for each of the three variables in LEADKIDS.

1) Compare the median of the exposed children with the maximum of the control children. What statement in the discussion does this confirm?

2) Compare the difference between the individual means of the exposed and control groups with the mean of the differences. On average, how much higher are the lead values for exposed children?

b. In contrast to part (a), notice that the difference between the individual medians of the exposed and control groups is *not* the same as the median for Diff. Why not? Which figure based on medians would you use if you were trying to give the most accurate view of the increase in lead exposure due to a parent working at the battery factory?

6.68 Analyzing Data, the Interpreting the Analyses, and Communicating the Results (4). A paired *t* test for the difference data in Exercise 6.67 is shown here.

```
Paired T for Exposed - Control

                N      Mean     StDev     SE Mean
Exposed        33     31.85     14.41        2.51
Control        33     15.88      4.54        0.79
Difference     33     15.97     15.86        2.76

95% CI for mean difference: (10.34, 21.59)
T-Test of mean difference = 0 (vs not = 0):
    T-Value = 5.78  P-Value = 0.000
```

The *p*-value in the output reads .000, which means that it is smaller than .0005 (1 chance in 2,000). Thus, it is extremely unlikely that we would see data as extreme as those actually collected unless workers at the battery factory were contaminating their children. We reject the null hypothesis and conclude that the difference between the lead levels of children in the exposed and control groups is large enough to be statistically significant.

The next question is whether the difference between the two groups is large enough to be of practical importance. This is a judgment for people who know about lead poisoning to make, not for statisticians. The best estimate of the true (population) mean difference is 15.97, or about 16. On average, children of workers in the battery plant have about 16 μg/dl more lead than their peers whose parents do not work in a lead-related industry. Almost any toxicologist would deem this increase to be dangerous and unacceptable. (The mean of the control group is also about 16. On average, the effect of having a parent who works in the battery factory is to double the lead level. Doubling the lead level brings the average value for exposed children to about 32, which is getting close to the level where medical treatment is required. Also remember that some toxicologists believe that any amount of lead is harmful to the neurological development of children.)

a. Should the *t* test we did have been one-sided? In practice, we must make the decision to do a one-sided test *before* the data are collected. We might argue that having a parent working at the battery factory could not *decrease* a child's exposure to lead.

1) Write the null hypothesis and its one-sided alternative in both words and symbols. Perform the test. How is its *p*-value related to the *p*-value for the two-sided test?

2) It might be tempting to argue that children of workers at a lead-using factory could not have generally lower levels of lead than children in the rest of the population. But can you imagine a scenario in which the mean levels would really be lower for exposed children?

b. We used a t test to confirm our impression that exposed children have more lead in their blood than their control counterparts. Although there is no clear reason to prefer nonparametric tests for these data, verify that they yield the same conclusion as the t test does.

Med. **6.69** The article ***"Increased Risk for Vitamin A Toxicity in Severe Hypertriglyceridemia" [Annals of Internal Medicine (1987) 105:877–879 (© American College of Physicians)]*** illustrates the importance of checking whether the appropriate conditions have been met prior to applying a statistical procedure. The data consist of the retinyl ester concentrations (mg/dl) of nine normal individuals and nine Type V hyperlipoproteinemic subjects.

Type V Subjects	1.4	2.5	4.6	0.0	0.0	2.9	1.9	4.0	2.0
Normal Subjects	30.9	134.6	13.6	28.9	434.1	101.7	85.1	26.5	44.8

a. Assess whether the data sets support the condition that both population distributions have normal distributions with equal variances.

b. Test for a difference in the mean retinyl ester concentrations of the two groups using the pooled t test, separate-variance t test, and Wilcoxon rank sum test. Use $\alpha = .01$.

c. Based on your conclusions in part (a), which test statistic would you recommend to test for a difference in the mean retinyl ester concentrations of the two groups?

CHAPTER 7

Inferences About Population Variances

7.1 Introduction and Abstract of Research Study

When people think of statistical inference, they usually think of inferences concerning population means. However, the population parameter that answers an experimenter's practical questions will vary from one situation to another. In many situations, the variability of a population's values is as important as the population mean. In the case of problems involving product improvement, product quality is defined as a product having mean value at the target value with low variability about the mean. For example, the producer of a drug product is certainly concerned with controlling the mean potency of tablets, but he or she must also worry about the variation in potency from one tablet to another. Excessive potency or an underdose could be very harmful to a patient. Hence, the manufacturer would like to produce tablets with the desired mean potency and with as little variation in potency (as measured by σ or σ^2) as possible. Another example is from the area of investment strategies. Investors search for a portfolio of stocks, bonds, real estate, and other investments having low risk. A measure used by investors to determine the uncertainty inherent in a particular portfolio is the variance in the value of the investments over a set period. At times, a portfolio with a high average value and a large standard deviation will have a value that is much lower than the average value. Investors thus need to examine the variability in the value of a portfolio along with its average value when determining its degree of risk.

Abstract of Research Study: Evaluation of Methods for Detecting *E. coli*

The outbreaks of bacterial disease in recent years due to the consumption of contaminated meat products have created a demand for new, rapid methods for detecting pathogens in meats that can be used in a meat surveillance program.

Under specific environmental conditions, certain strains of bacteria such as *E. coli* are capable of causing hemorrhagic colitis, hemolytic uremic syndrome, and even death. An effective pathogen surveillance program requires three main attributes: (1) a probability-based sampling plan (as described in Chapter 2), (2) a method capable of efficiently removing viable organisms from the target surface of animals, and (3) a repeatable, accurate, and practical microbial test for the target pathogen. The paper ***"Repeatability of the Petrifilm HEC Test and Agreement with a Hydrophobic Grid Membrane Filtration Method for the Enumeration of* Escherichia coli *0157:H7 on Beef Carcasses" (Power et al., 1998)***, describes a formal comparison between a new microbial method for the detection of *E. coli,* the Petrifilm HEC test, and an elaborate laboratory-based procedure, hydrophobic grid membrane filtration (HGMF). The HEC test is easier to inoculate, more compact to incubate, and safer to handle than conventional procedures. However, it was necessary to compare the performance of the HEC test to that of the HGMF procedure in order to determine if the HEC test might be a viable method for detecting *E. coli*.

What aspects of the *E. coli* counts obtained by HEC and HGMF should be of interest to the researchers? A comparison of just the mean concentrations obtained by the two procedures would indicate whether or not the two procedures were in agreement with respect to the average readings over a large number of determinations. However, we would not know if HEC was more variable in its determination of *E. coli* than HGMF. For example, consider the two distributions in Figure 7.1. Suppose the distributions represent the population of *E. coli* concentration determinations from HEC and HGMF for a situation in which the true *E. coli* concentration is 7 $\log_{10}$ CFU/ml. The distributions would indicate that the HEC evaluation of a given meat sample may yield a reading very different from the true *E. coli* concentration, whereas the individual readings from HGMF are more likely to be near the true concentration. In this type of situation, it is crucial to compare both the means and the standard deviations of the two procedures. In fact, we need to examine other aspects of the relationship between HEC and HGMF determinations in order to evaluate the comparability of the two procedures.

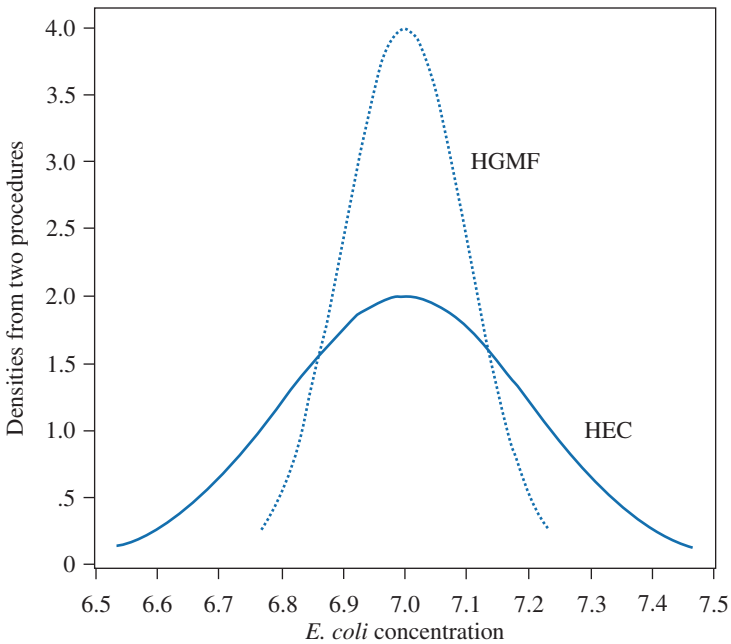


FIGURE 7.1
Hypothetical distribution of *E. coli* concentrations from HEC and HGMF

The experiment was designed to have two phases. Phase One of the study was to apply both procedures to pure cultures of *E. coli* representing 10^7 CFU/ml of strain E318N. Based on the specified degree of precision in estimating the *E. coli* level, it was determined that the HEC and HGMF procedures would be applied to 24 pure cultures each (we will discuss how the sample size of 24 was selected later in this chapter). Phase Two of the study was to apply both procedures to artificially contaminated beef. Portions of beef trim were obtained from three Holstein cows that had tested negatively for *E. coli*. Eighteen portions of beef trim were obtained from the cows and then contaminated with *E. coli*. The HEC and HGMF procedures were then applied to a portion of each of the 18 samples. The two procedures yielded *E. coli* concentrations ($\log_{10}$ CFU/ml). The data in this case would be 18 paired samples. The researchers were interested in determining a model to relate the two procedures' determinations of *E. coli* concentrations. We will consider only Phase One in this chapter. We will consider Phase Two in Chapter 11 in our development of model building and calibration. The researchers found that the HEC test showed excellent repeatability and excellent agreement with the HGMF method. In a later section of this chapter and in Chapter 11, we will demonstrate how the researchers reached these conclusions.

Inferential problems about population variances are similar to the problems addressed in making inferences about population means. We must construct point estimators, confidence intervals, and test statistics from the randomly sampled data to make inferences about the variability in the population values. We then can state our degree of certainty that observed differences in the sample data convey differences in the population parameters.

7.2 Estimation and Tests for a Population Variance

The sample variance

$$s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n - 1}$$

can be used for inferences concerning a population variance σ^2. For a random sample of n measurements drawn from a population with mean μ and variance σ^2, s^2 is an **unbiased estimator** of σ^2. If the population distribution is normal, then the sampling distribution of s^2 can be specified as follows. From repeated samples of size n from a normal population whose variance is σ^2, calculate the statistic $(n - 1)s^2/\sigma^2$, and plot the histogram for these values. The shape of the histogram is similar to those depicted in Figure 7.2 because it can be shown that the statistic $(n - 1)s^2/\sigma^2$ follows a **chi-square distribution with df = $n - 1$**. The mathematical formula for the chi-square (χ^2, where χ is the Greek letter chi) probability distribution is very complex, so we will not display it. However, some of the properties of the distribution are as follows:

unbiased estimator

chi-square distribution with df = $n - 1$

1. The chi-square distribution is positively skewed with values between 0 and ∞ (see Figure 7.2).
2. There are many chi-square distributions, and they are labeled by the parameter degrees of freedom (df). Three such chi-square distributions are shown in Figure 7.2 with df = 5, 15, and 30, respectively.
3. The mean and variance of the chi-square distribution are given by $\mu = \text{df}$ and $\sigma^2 = 2\text{df}$. For example, if the chi-square distribution has df = 30, then the mean and variance of that distribution are $\mu = 30$ and $\sigma^2 = 60$.

FIGURE 7.2
Densities of the chi-square (df = 5, 15, 30) distribution

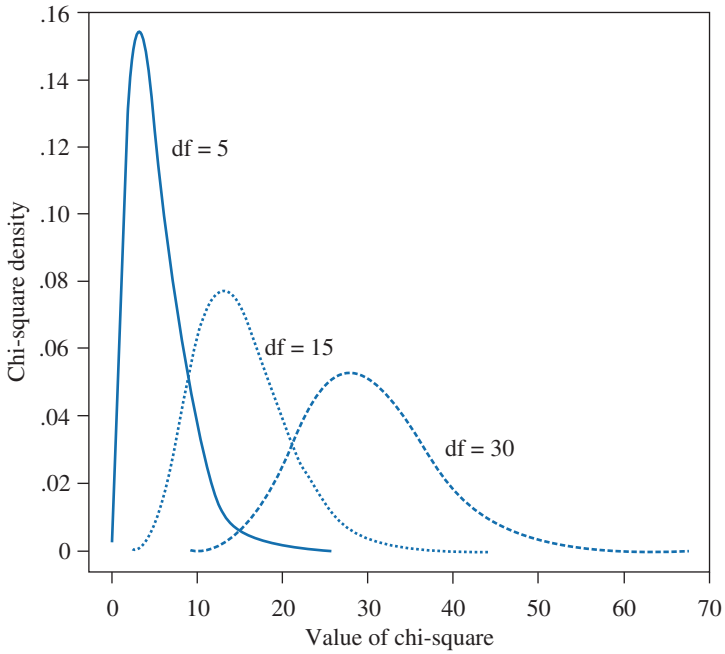
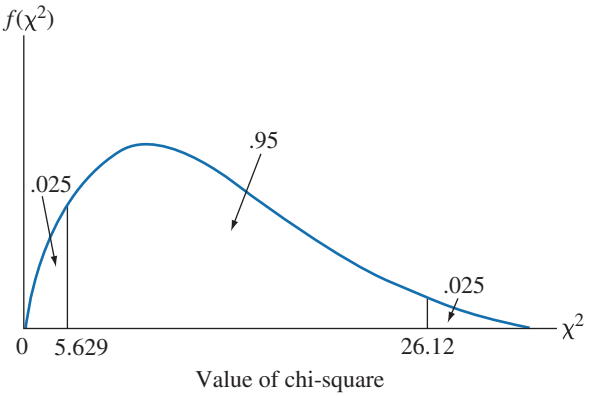
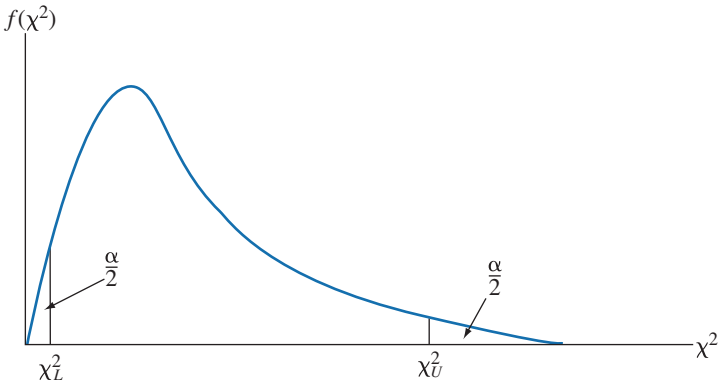


FIGURE 7.3
Critical values of the chi-square distribution with df = 14



Upper-tail values of the chi-square distribution can be found in Table 7 in the Appendix. Entries in the table are values of χ^2 that have an area α to the right under the curve. The degrees of freedom are specified in the left column of the table, and values of α are listed across the top of the table. Thus, for df = 14, the value of chi-square with an area $\alpha = .025$ to its right under the curve is 26.12 (see Figure 7.3). To determine the value of chi-square with an area of .025 to its left under the curve, we compute $\alpha = 1 - .025$ and obtain 5.629 from Table 7 in the Appendix. Combining these two values, we have that the area under the curve between 5.629 and 26.12 is $1 - .025 - .025 = .95$. (See Figure 7.3.) We can use this information to form a confidence interval for σ^2. Because the chi-square distribution is not symmetrical, the confidence intervals based on this distribution do not have the usual form, estimate $\pm$ error, as we saw for μ and $\mu_1 - \mu_2$. The $100(1 - \alpha)\%$ confidence interval for σ^2 is obtained by dividing the estimator of σ^2, s^2, by the lower and upper $\alpha/2$ percentiles, χ^2_L and χ^2_U, as described next.

FIGURE 7.4
Upper-tail and lower-tail values of chi-square



General Confidence Interval for σ^2 (or σ) with Confidence Coefficient $(1 - \alpha)$

$$\left(\frac{(n-1)s^2}{\chi^2_U}, \frac{(n-1)s^2}{\chi^2_L}\right)$$

where χ^2_U is the upper-tail value of chi-square for df $= n - 1$ with area $\alpha/2$ to its right and χ^2_L is the lower-tail value with area $\alpha/2$ to its left (see Figure 7.4). We can determine χ^2_U and χ^2_L for a specific value of df by obtaining the critical values in Table 7 of the Appendix corresponding to $\alpha/2$ and $1 - \alpha/2$, respectively. (*Note:* The confidence interval for σ is found by taking square roots throughout.)

The upper and lower α percentiles of the chi-square distribution can be obtained using the R function ***qchisq*(α, *df*)**:

The upper $\alpha/2$ percentile is given by $\chi^2_U =$ ***qchisq*(1 − α/2, *df*)**.
The lower $\alpha/2$ percentile is given by $\chi^2_L =$ ***qchisq*(α/2, *df*)**.

EXAMPLE 7.1

The machine that fills 500-gram coffee containers for a large food processor is monitored by the quality control department. Ideally, the amount of coffee in a container should vary only slightly about the nominal 500-gram value. If the variation was large, then a large proportion of the containers would be either underfilled, thus cheating the customer, or overfilled, thus resulting in economic loss to the company. The machine was designed so that the weights of the 500-gram containers would have a normal distribution with a mean value of 506.6 grams and a standard deviation of 4 grams. This would produce a population of containers in which at most 5% of the containers weighed less than 500 grams. To maintain a population in which at most 5% of the containers are underweight, a random sample of 30 containers is selected every hour to be weighed. These data are then used to determine whether the mean and standard deviation are maintained at their nominal values. The weights from one of the hourly samples are given here:

501.4	498.0	498.6	499.2	495.2	501.4	509.5	494.9	498.6	497.6
505.5	505.1	499.8	502.4	497.0	504.3	499.7	497.9	496.5	498.9
504.9	503.2	503.0	502.6	496.8	498.2	500.1	497.9	502.2	503.2

Estimate the mean and standard deviation of the weights of the 30 coffee containers using a 99% confidence interval.

Solution For these data, we find

$$\bar{y} = 500.453 \quad \text{and} \quad s = 3.433$$

To use our method for constructing a confidence interval for μ and σ, we must first check whether the weights are a random sample from a normal population. Figure 7.5 is a normal probability plot of the 30 weights. The 30 values fall near the straight line. Thus, the normality condition appears to be satisfied. The confidence coefficient for this example is $1 - \alpha = .99$. The upper-tail chi-square value can be obtained from Table 7 in the Appendix for df $= n - 1 = 29$ and $\alpha/2 = .005$. Similarly, the lower-tail chi-square value is obtained from Table 7 with $1 - \alpha/2 = .995$. Thus,

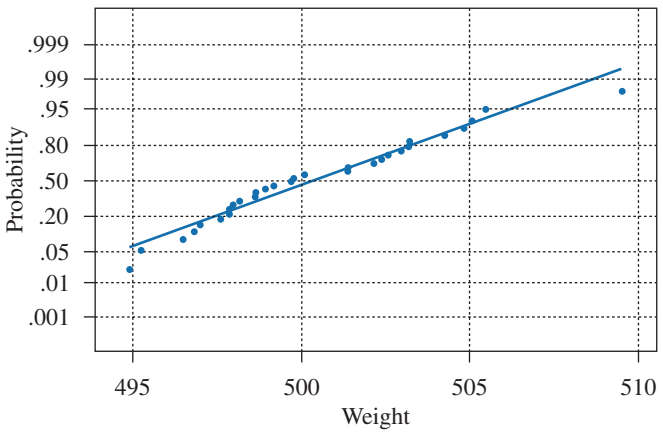
$$\chi^2_L = 13.12 \quad \text{and} \quad \chi^2_U = 52.34$$

Using the R function, the upper $.01/2 = .005$ and lower $.01/2 = .005$ percentiles for a chi-square distribution with df $= 29$ are obtained as follows:

$$\chi^2_U = \textbf{\textit{qchisq}}(1 - .01/2, 29) = \textbf{\textit{qchisq}}(.995, 29) = 52.34$$

$$\chi^2_L = \textbf{\textit{qchisq}}(.01/2, 29) = \textbf{\textit{qchisq}}(.005, 29) = 13.12$$

FIGURE 7.5
Normal probability plot of container weights



The 99% confidence interval for σ is then

$$\left(\sqrt{\frac{29(3.433)^2}{52.34}}, \sqrt{\frac{29(3.433)^2}{13.12}}\right) = (2.56, 5.10)$$

Thus, we are 99% confident that the standard deviation in the weights of the coffee containers lies between 2.56 and 5.10 grams. The designed value for σ, 4 grams, falls within our confidence interval. Using our results from Chapter 5, a 99% confidence interval for μ is

$$500.453 \pm 2.756 \frac{3.433}{\sqrt{30}} = 500.453 \pm 1.73 = (498.7, 502.2)$$

Thus, it appears the machine is underfilling the containers because 506.6 grams does not fall within the confidence limits. ■

In addition to estimating a population variance, we can construct a statistical test of the null hypothesis that σ^2 equals a specified value, σ_0^2. This test procedure is summarized next.

Statistical Test for σ^2 (or σ)

H_0: **1.** $\sigma^2 \le \sigma_0^2$ $\quad$ H_a: **1.** $\sigma^2 > \sigma_0^2$

2. $\sigma^2 \ge \sigma_0^2$ $\quad$ **2.** $\sigma^2 < \sigma_0^2$

3. $\sigma^2 = \sigma_0^2$ $\quad$ **3.** $\sigma^2 \ne \sigma_0^2$

T.S.: $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$

R.R.: For a specified value of α,

1. Reject H_0 if χ^2 is greater than χ_U^2, the upper-tail value for α and df $= n - 1$.
2. Reject H_0 if χ^2 is less than χ_L^2, the lower-tail value for $1 - \alpha$ and df $= n - 1$.
3. Reject H_0 if χ^2 is greater than χ_U^2, based on $\alpha/2$ and df $= n - 1$, or less than χ_L^2, based on $1 - \alpha/2$ and df $= n - 1$.

Check assumptions and draw conclusions.

EXAMPLE 7.2

New guidelines define persons as diabetic if results from their fasting plasma glucose tests on two different days are 126 milligrams per deciliter (mg/dL) or higher. People who have a reading of between 110 and 125 are considered in danger of becoming diabetic, as their ability to process glucose is impaired. These people should be tested more frequently and counseled about ways to lower their blood sugar level and reduce the risk of heart disease.

Amid sweeping changes in U.S. health care, the trend toward cost-effective self-care products used in the home emphasizes prevention and early intervention. The home test kit market is offering faster-acting and easier-to-use products that lend themselves to being used in less-sophisticated environments to meet consumers' needs. A home blood sugar (glucose) test measures the level of glucose in your blood at the time of testing. The test can be done at home, or anywhere, using a small portable machine called a blood glucose meter. People who take insulin to control their diabetes may need to check their blood glucose level several times a day. Testing blood sugar at home is often called home blood sugar monitoring or self-testing.

Home glucose meters are not usually as accurate as laboratory measurement. Problems arise when the machines are not properly maintained and, more importantly, when the persons conducting the tests are the patients themselves, who may be quite elderly and in poor health. In order to evaluate the variability in readings from such devices, blood samples with a glucose level of 200 mg/dL are given to 20 diabetic patients to perform a self-test for glucose level. Trained technicians using the same self-test equipment obtain readings that have a standard deviation of 5 mg/dL. The manufacturer of the equipment claims that, with minimal instruction, anyone can obtain the same level of

consistency in their measurements. The readings from the 20 diabetic patients are given here:

203.1	184.5	206.8	211.0	218.3	174.2	193.2	201.9	199.9	194.3
199.4	193.6	194.6	187.2	197.8	184.3	196.1	196.4	197.5	187.9

Use these data to determine whether there is sufficient evidence that the variability in readings from the diabetic patients is higher than the manufacturer's claim. Use $\alpha = .05$.

Solution The manufacturer claims that the diabetic patients should have a standard deviation of 5 mg/dL. The appropriate hypotheses are

H_0: $\sigma^2 \leq 5$ (manufacturer's claim is correct)

H_a: $\sigma^2 > 5$ (manufacturer's claim is false)

In order to apply our test statistic to these hypotheses, it is necessary to check whether the data appear to have been generated from a normally distributed population. From Figure 7.6, we observe that the plotted points fall relatively close to the straight line and that the p-value for testing normality is greater than .10. Thus, the normality condition appears to be satisfied. From the 20 data values, we compute the sample standard deviation, $s = 9.908$. The test statistic and rejection regions are as follows:

$$\text{T.S.:} \quad \chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{19(9.908)^2}{(5)^2} = 74.61$$

R.R.: For $\alpha = .05$, the null hypothesis, H_0, is rejected if the value of the T.S. is greater than 30.14, obtained from Table 7 in the Appendix for $\alpha = .05$ and df $= n - 1 = 19$.

Conclusion: Since the computed value of the T.S., 74.61, is greater than the critical value of 30.14, there is sufficient evidence to reject H_0, the manufacturer's claim, at the .05 level. In fact, the p-value of the T.S. is p-value $= P(\chi^2_{19} \geq 74.61) < P(\chi^2_{19} \geq 43.82) = .001$ using Table 7 from the Appendix.

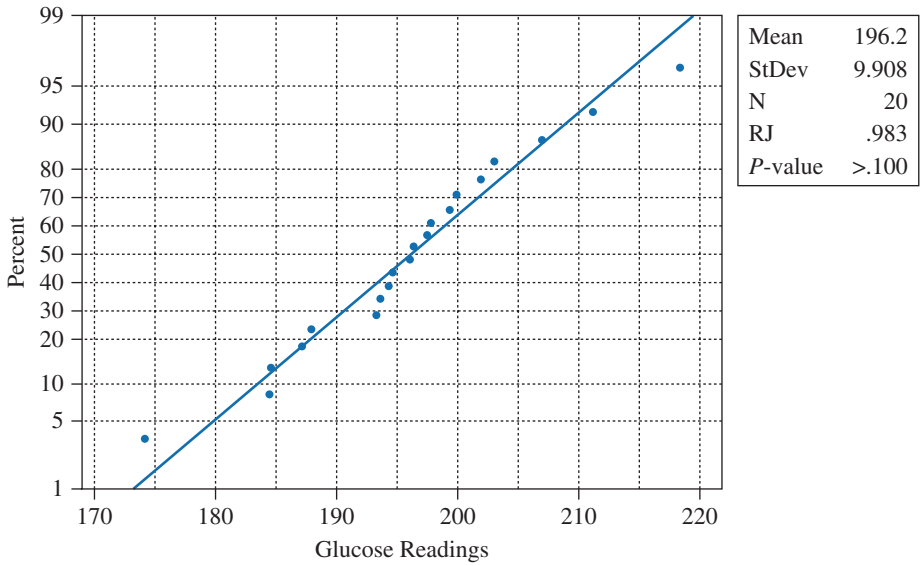


FIGURE 7.6
Normal probability plot for glucose readings

Using the R function ***pchisq***$(y,\ \text{df}) = P(\chi^2 \le y)$, the p-value is computed to be $p\text{-value} = P(\chi^2 \ge 74.61) = 1 - $ ***pchisq***$(74.61, 19) = .00000002$. Thus, there is very strong evidence that patients using the self-test for glucose may have larger variability in their readings than what the manufacturer claimed. In fact, to further assess the size of this standard deviation, a 95% confidence interval for σ is given by

$$\left(\sqrt{\frac{(20-1)(9.908)^2}{32.85}}, \sqrt{\frac{(20-1)(9.908)^2}{8.907}}\right) = (7.53, 14.47)$$

Therefore, the standard deviation in glucose measurements for the diabetic patients is potentially considerably higher than the standard deviation for the trained technicians. ■

The inference methods about σ are based on the condition that the random sample is selected from a population having a normal distribution similar to the requirements for using t distribution–based inference procedures. However, when sample sizes are moderate to large, the t distribution–based procedures can be used to make inferences about μ even when the normality condition does not hold because for moderate to large sample sizes the Central Limit Theorem provides that the sampling distribution of the sample mean is approximately normal. Unfortunately, the same type of result does not hold for the chi-square–based procedures for making inferences about σ; that is, if the population distribution is distinctly nonnormal, then these procedures for σ are not appropriate even if the sample size is large. Population nonnormality, in the form of skewness or heavy tailedness, can have serious effects on the nominal significance and confidence probabilities for σ. If a boxplot or normal probability plot of the sample data shows substantial skewness or a substantial number of outliers, the chi-square–based inference procedures should not be applied. There are some alternative approaches that involve computationally elaborate inference procedures. One such procedure is the bootstrap. Bootstrapping is a technique that provides a simple and practical way to estimate the uncertainty in sample statistics like the sample variance. We can use bootstrap techniques to estimate the sampling distribution of a sample variance. The estimated sampling distribution is then manipulated to produce confidence intervals for σ and rejection regions for tests of hypotheses about σ. Information about bootstrapping can be found in *An Introduction to the Bootstrap* (Efron and Tibshirani, 1993) and *Randomization, Bootstrap and Monte Carlo Methods in Biology* (Manly, 1998).

EXAMPLE 7.3

A simulation study was conducted to investigate the effect on the level of the chi-square test of sampling from heavy-tailed and skewed distributions rather than the required normal distribution. The five distributions were normal, uniform (short-tailed), t distribution with df = 5 (heavy-tailed), and two gamma distributions, one slightly skewed and the other heavily skewed. Some summary statistics about the distributions are given in Table 7.1.

Note that each of the distributions has the same variance, $\sigma^2 = 100$, but the skewness and kurtosis of the distributions vary. Skewness is a measure of lack of symmetry, and kurtosis is a measure of the peakedness or flatness of a distribution. From each of the distributions, 2,500 random samples of sizes 10, 20,

TABLE 7.1
Summary statistics for distributions in simulation

Summary Statistic	Distribution				
	Normal	Uniform	t (df = 5)	Gamma (shape = 1)	Gamma (shape = .1)
Mean	0	17.32	0	10	3.162
Variance	100	100	100	100	100
Skewness	0	0	0	2	6.32
Kurtosis	3	1.8	9	9	63

and 50 were selected, and a test of H_0: $\sigma^2 \leq 100$ versus H_a: $\sigma^2 > 100$ and a test of H_0: $\sigma^2 \geq 100$ versus H_a: $\sigma^2 < 100$ were conducted using $\alpha = .05$ for both sets of hypotheses. A chi-square test of variance was performed for each of the 2,500 samples of the various sample sizes from each of the five distributions. The results are given in Table 7.2. What do the results indicate about the sensitivity of the test to sampling from a nonnormal population?

TABLE 7.2
Proportion of times H_0 was rejected ($\alpha = .05$)

Sample Size	H_α: $\sigma^2 > 100$				
	Distribution				
	Normal	Uniform	t	Gamma (1)	Gamma (.1)
$n = 10$	.047	.004	.083	.134	.139
$n = 20$	.052	.006	.103	.139	.175
$n = 50$	.049	.004	.122	.156	.226
Sample Size	H_α: $\sigma^2 < 100$				
	Distribution				
	Normal	Uniform	t	Gamma (1)	Gamma (.1)
$n = 10$	.046	.018	.119	.202	.213
$n = 20$	.050	.011	.140	.213	.578
$n = 50$	.051	.018	.157	.220	.528

Solution The values in Table 7.2 are estimates of the probability of a Type I error, α, for the chi-square test about variances. When the samples are taken from a normal population, the actual probabilities of a Type I error are very nearly equal to the nominal $\alpha = .05$ value. When the population distribution is symmetric with shorter tails than a normal distribution, the actual probabilities are smaller than .05, whereas for a symmetric distribution with heavy tails, the Type I error probabilities are much greater than .05. Also, for the two skewed distributions, the actual α values are much larger than the nominal .05 value. Furthermore, as the population distribution becomes more skewed, the deviation from .05 increases. From these results, there is strong evidence that the claimed α value of the chi-square test of a population variance is very sensitive to nonnormality. *This strongly reinforces our recommendation to evaluate the normality of the data prior to conducting the chi-square test of a population variance.* ■

7.3 Estimation and Tests for Comparing Two Population Variances

In the research study about *E. coli* detection methods, we are concerned about comparing the standard deviations of the two procedures. In many situations in which we are comparing two processes or two suppliers of a product, we need to compare the standard deviations of the populations associated with process measurements. The test developed in this section requires that the two population distributions both have normal distributions. We are interested in comparing the variance of population 1, σ_1^2, to the variance of population 2, σ_2^2.

When random samples of sizes n_1 and n_2 have been independently drawn from two normally distributed populations, the ratio

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2}$$

***F* distribution**

possesses a probability distribution in repeated sampling referred to as an ***F* distribution**. The formula for the probability distribution is omitted here, but we will specify its properties.

Properties of the *F* Distribution

1. Unlike t or z but like χ^2, F can assume only positive values.
2. The F distribution, unlike the normal distribution or the t distribution but like the χ^2 distribution, is nonsymmetrical. (See Figure 7.7.)
3. There are many F distributions, and each one has a different shape. We specify a particular one by designating the degrees of freedom associated with s_1^2 and s_2^2. We denote these quantities by df_1 and df_2, respectively. (See Figure 7.7.)
4. Tail values for the F distribution are tabulated and appear in Table 8 in the Appendix.

FIGURE 7.7
Densities of two F distributions

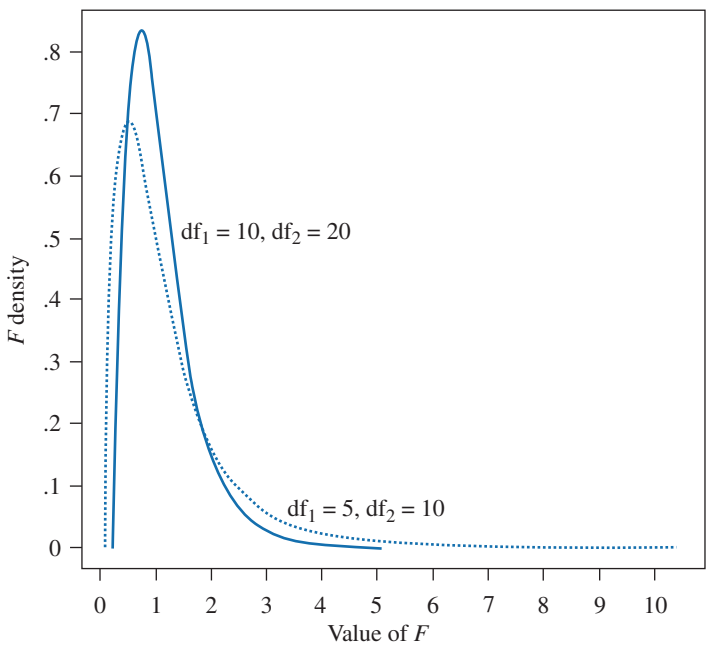


FIGURE 7.8
Critical value for the F distributions ($df_1 = 5, df_2 = 10$)

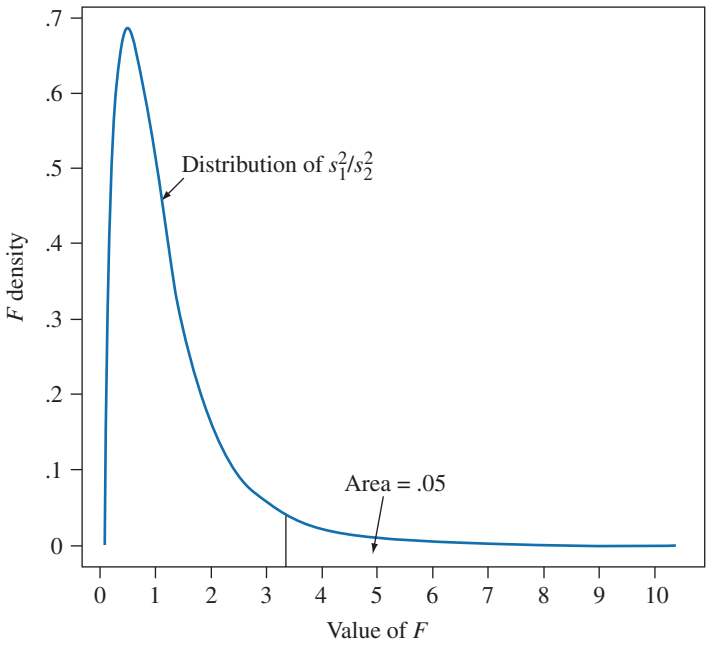


Table 8 in the Appendix records upper-tail values of F corresponding to areas $\alpha = .25, .10, .05, .025, .01, .005$, and $.001$. The degrees of freedom for s_1^2, designated by df_1, are indicated across the top of the table; df_2, the degrees of freedom for s_2^2, appear in the first column to the left. Values of α are given in the next column. Thus, for $df_1 = 5$ and $df_2 = 10$, the critical values of F corresponding to $\alpha = .25, .10, .05, .025, .01, .005$, and $.001$ are, respectively, 1.59, 2.52, 3.33, 4.24, 5.64, 6.78, and 10.48. It follows that only 5% of the measurements from an F distribution with $df_1 = 5$ and $df_2 = 10$ would exceed 3.33 in repeated sampling. (See Figure 7.8.) Similarly, for $df_1 = 24$ and $df_2 = 10$, the critical values of F corresponding to tail areas of $\alpha = .01$ and $.001$ are, respectively, 4.33 and 7.64.

A statistical test comparing σ_1^2 and σ_2^2 utilizes the test statistic s_1^2/s_2^2. When $\sigma_1^2 = \sigma_2^2$, $\sigma_1^2/\sigma_2^2 = 1$ and s_1^2/s_2^2 follows an F distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$. For a one-tailed alternative hypothesis, the designation of which population is 1 and which population is 2 is made such that H_a is of the form $\sigma_1^2 > \sigma_2^2$. Then the rejection region is located in the upper tail of the F distribution.

We summarize the test procedure next.

A Statistical Test Comparing σ_1^2 and σ_2^2

H_0: **1.** $\sigma_1^2 \le \sigma_2^2$ $\quad$ H_a: **1.** $\sigma_1^2 > \sigma_2^2$
2. $\sigma_1^2 = \sigma_2^2$ $\quad$ **2.** $\sigma_1^2 \ne \sigma_2^2$

T.S.: $F = s_1^2/s_2^2$

R.R.: For a specified value of α with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$,

1. Reject H_0 if $F \ge F_{\alpha, df_1, df_2}$.

2. Reject H_0 if $F \le F_{1-\alpha/2, df_1, df_2}$ or if $F \ge F_{\alpha/2, df_1, df_2}$.

Table 8 in the Appendix provides the upper percentiles of the F distribution. The lower percentiles are obtained from the upper percentiles using the following relationship. Let F_{α, df_1, df_2} be the upper α percentile and $F_{1-\alpha, df_1, df_2}$ be the lower α percentile of an F distribution with df_1 and df_2. Then

$$F_{1-\alpha, df_1, df_2} = \frac{1}{F_{\alpha, df_2, df_1}}$$

Note that the degrees of freedom have been reversed for the upper F percentile on the right-hand side of the equation.

The upper and lower α percentiles of the F distribution can be obtained using the R function ***qf***(α, df_1, df_2).

The upper α percentile is given by $F_{\alpha, df_1, df_2} =$ ***qf*****$(1-\alpha, df_1, df_2)$**

The lower α percentile is given by $F_{1-\alpha, df_1, df_2} =$ ***qf*****(α, df_1, df_2)**

EXAMPLE 7.4

Determine the lower .025 percentile for an F distribution with $df_1 = 7$ and $df_2 = 10$.

Solution From Table 8 in the Appendix, the upper .025 percentile for the F distribution with $df_1 = 10$ and $df_2 = 7$ is $F_{.025, 10, 7} = 4.76$. Thus, the lower .025 percentile is given by

$$F_{.975, 7, 10} = \frac{1}{F_{.025, 10, 7}} = \frac{1}{4.76} = 0.21$$

Using the R function, the upper .025 percentile for an F distribution with $df_1 = 10$ and $df_2 = 7$ is obtained as follows:

$$F_{.025, 10, 7} = \textbf{\textit{qf}}(1 - .025, 10, 7) = 4.7611$$

Similarly, the lower .025 percentile for an F distribution with $df_1 = 7$ and $df_2 = 10$ is given by

$$F_{.975, 7, 10} = \textbf{\textit{qf}}(.025, 7, 10) = .2100 \blacksquare$$

EXAMPLE 7.5

In the research study discussed in Chapter 6, we were concerned with assessing the restoration of land damaged by an oil spill. Random samples of 80 tracts from the unaffected and oil-spill areas were selected for use in the assessment of how well the oil-spill area was restored to its prespill status. Measurements of flora density were taken on each of the 80 tracts. These 80 densities were then used to test whether the unaffected (control) tracts had a higher mean density than the restored spill sites: H_a: $\mu_{Con} > \mu_{Spill}$. A confidence interval was also placed on the effect size: $\mu_{Con} - \mu_{Spill}$.

We mentioned in Chapter 6 that in selecting the test statistic and constructing confidence intervals for $\mu_1 - \mu_2$ we require that the random samples be drawn from normal populations that may have different means *but* that must have equal variances in order to apply the *pooled* t *procedures*. Use the sample data summarized next to test the equality of the population variances for the flora densities. Use $\alpha = .05$.

Control plots: $\bar{y}_1 = 38.48$ $s_1 = 16.37$ $n_1 = 40$

Spill plots: $\bar{y}_2 = 26.93$ $s_2 = 9.88$ $n_2 = 40$

Solution The four parts of the statistical test of H_0: $\sigma_1^2 = \sigma_2^2$ follow:

$$H_0: \ \sigma_1^2 = \sigma_2^2$$
$$H_a: \ \sigma_1^2 \neq \sigma_2^2$$
$$\text{T.S.:} \ F = \frac{s_1^2}{s_2^2} = \frac{(16.37)^2}{(9.88)^2} = 2.75$$

Prior to setting the rejection region, we must first determine whether the two random samples appear to be from normally distributed populations. Figures 6.9 and 6.10(a) and (b) indicate that the oil-spill sites appear to be selected from a normal distribution. However, the control sites appear to have a distribution somewhat skewed to the left. Although the normality condition is not exactly satisfied, we will still apply the F test to this situation. In Section 7.4, we will introduce a test statistic that is not as sensitive to deviations from normality.

R.R.: For a two-tailed test with $\alpha = .05$, we reject H_0 if $F \geq F_{.025,39,39} \approx 1.88$ or if $F \leq F_{.975,39,39} \approx 1/1.88 = .53$ (we used the values for $df_1 = df_2 = 40$ as an approximation, since Table 8 in the Appendix does not have values for $df_1 = df_2 = 39$). Using the R function, the actual values are 1.8907 and .5289.

Conclusion: Because $F = 2.75$ exceeds 1.88, we reject H_0: $\sigma_1^2 = \sigma_2^2$ and conclude that the two populations have unequal variances. Thus, our decision to use the separate-variance t test in the analysis of the oil-spill data was the correct decision. ■

In Chapter 6, our tests of hypotheses concerned either population means or a shift parameter. For both types of parameters, it was important to provide an estimate of the *effect size* along with the conclusion of the test of hypotheses. In the case of testing population means, the effect size was stated in terms of the difference in the two means: $\mu_1 - \mu_2$. When comparing population variances, the appropriate measure is the ratio of the population variances: σ_1^2/σ_2^2. Thus, we need to formulate a confidence interval for the ratio σ_1^2/σ_2^2. A $100(1-\alpha)\%$ confidence interval for this ratio is given here.

General Confidence Interval for σ_1^2/σ_2^2 with Confidence Coefficient $(1-\alpha)$

$$\left(\frac{s_1^2}{s_2^2}F_L, \frac{s_1^2}{s_2^2}F_U\right)$$

where $F_U = F_{\alpha/2, df_2, df_1}$ and $F_L = F_{1-\alpha/2, df_2, df_1} = 1/F_{\alpha/2, df_1, df_2}$ with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$. (*Note:* A confidence interval for σ_1/σ_2 is found by taking the square root of the endpoints of the confidence interval for σ_1^2/σ_2^2.)

EXAMPLE 7.6

Refer to Example 7.5. We rejected the hypothesis that the variances of flora density for the control and oil-spill sites were equal. The researchers would then want to estimate the magnitude of the disagreement in the variances. Using the data in Example 7.5, construct a 95% confidence interval for σ_1^2/σ_2^2.

Solution The confidence interval for the ratio of the two variances is given by

$$\left(\frac{s_1^2}{s_2^2}F_L, \frac{s_1^2}{s_2^2}F_U\right)$$

where $F_L = F_{1-\alpha/2, n_2-1, n_1-1} = F_{.975,39,39} = .53$ and $F_U = F_{\alpha/2, n_2-1, n_1-1} = F_{.025,39,39} = 1.89$. Thus, we have the 95% confidence interval given by

$$\left(\frac{(16.37)^2}{(9.88)^2}.53, \frac{(16.37)^2}{(9.88)^2}1.89\right) = (1.45, 5.19)$$

Thus, we are 95% confident that the flora density in the control plots is between 1.45 and 5.19 times as variable as the oil spill plots. ■

It should be noted that although our estimation procedure for σ_1^2/σ_2^2 is appropriate for any confidence coefficient $(1 - \alpha)$, Table 8 in the Appendix allows us to construct confidence intervals for σ_1^2/σ_2^2 with the more commonly used confidence coefficients, such as .90, .95, .98, .99, and so on. For more detailed tables of the F distribution, see ***Pearson and Hartley (1966) or use the R function qf***.

EXAMPLE 7.7

The life length of an electrical component was studied under two operating voltages, 110 and 220. Ten different components were randomly assigned to operate at 110 volts, and 16 different components were randomly assigned to operate at 220 volts. The times to failure (in hundreds of hours) for the 26 components were obtained and yielded the following summary statistics and normal probability plots (see Figures 7.9 and 7.10 as well as Table 7.3).

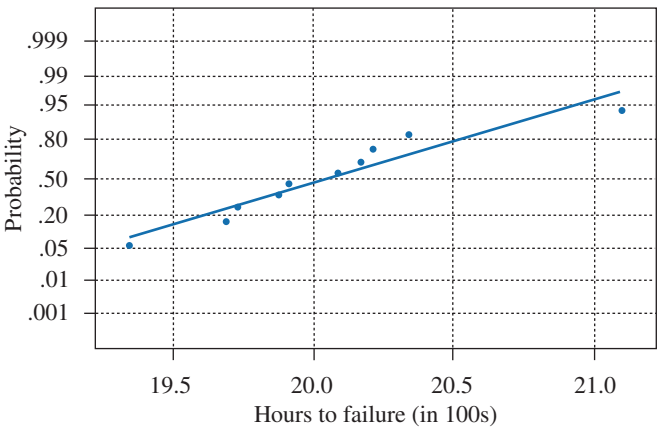


FIGURE 7.9
Normal probability plot for life length under 110 volts

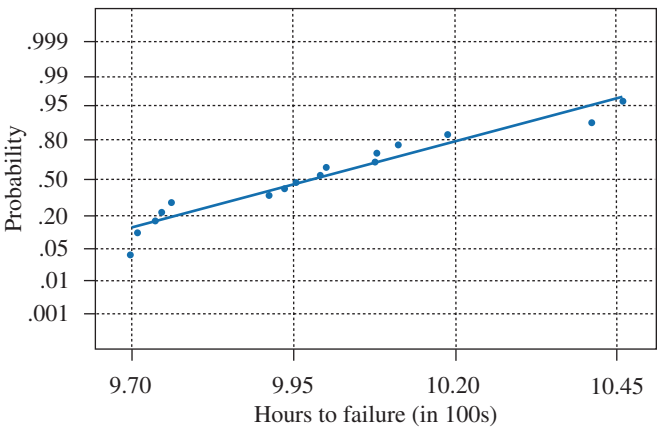


FIGURE 7.10
Normal probability plot for life length under 220 volts

TABLE 7.3
Life length summary statistics

Voltage	Sample Size	Mean	Standard Deviation
110	10	20.04	.474
220	16	9.99	.233

The researchers wanted to estimate the relative size of the variation in life length under 110 and 220 volts. Use the data to construct a 90% confidence interval for σ_1/σ_2, the ratio of the standard deviations in life lengths for the components under the two operating voltages.

Solution Before constructing the confidence interval, it is necessary to check whether the two populations of life lengths were both normally distributed. From the normal probability plots, it would appear that both samples of life lengths are from normal distributions. Next, we need to find the upper and lower $\alpha/2 = .10/2 = .05$ percentiles for the F distribution with $df_1 = 10 - 1 = 9$ and $df_2 = 16 - 1 = 15$. From Table 8 in the Appendix, we find

$$F_U = F_{.05,15,9} = 3.01 \quad \text{and} \quad F_L = F_{.95,15,9} = 1/F_{.05,9,15} = 1/2.59 = .386$$

Substituting into the confidence interval formula, we have a 90% confidence interval for σ_1^2/σ_2^2:

$$\frac{(.474)^2}{(.233)^2}.386 \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{(.474)^2}{(.233)^2}3.01$$

$$1.5975 \le \frac{\sigma_1^2}{\sigma_2^2} \le 12.4569$$

It follows that our 90% confidence interval for σ_1/σ_2 is given by

$$\sqrt{1.5975} \le \frac{\sigma_1}{\sigma_2} \le \sqrt{12.4569} \quad \text{or} \quad 1.26 \le \frac{\sigma_1}{\sigma_2} \le 3.53$$

Thus, we are 90% confident that σ_1 is between 1.26 and 3.53 times as large as σ_2. ■

A simulation study was conducted to investigate the effect on the level of the F test of sampling from heavy-tailed and skewed distributions rather than the required normal distribution. The five distributions were described in Example 7.3.

For each pair of sample sizes $(n_1, n_2) = (10, 10)$, $(10, 20)$, or $(20, 20)$, random samples of the specified sizes were selected from one of the five distributions. A test of H_0: $\sigma_1^2 = \sigma_2^2$ versus H_a: $\sigma_1^2 \neq \sigma_2^2$ was conducted using an F test with $\alpha = .05$. This process was repeated 2,500 times for each of the five distributions and three sets of sample sizes. The results are given in Table 7.4.

The values given in Table 7.4 are estimates of the probability of Type I errors, α, for the F test of equality of two population variances. When the samples are from a normally distributed population, the value of α is nearly equal to the nominal level of .05 for all three pairs of sample sizes. This is to be expected because the F test was constructed to test hypotheses when the population distributions have normal distributions. However, when the population distribution is a symmetric short-tailed distribution like the uniform distribution, the value of α is much smaller than the specified value of .05. Thus, the probability of Type II errors for

the F test would most likely be much larger than what would occur when sampling from normally distributed populations. When we have population distributions that are symmetric and heavy-tailed, like the t with df = 5, the values of α are two to three times larger than the specified value of .05. Thus, the F test commits many more Type I errors than would be expected when the population distributions are of this type. A similar problem occurs when we sample with skewed population distributions such as the two gamma distributions. In fact, the Type I error rates are extremely large in these situations, thus rendering the F test invalid for these types of distributions.

TABLE 7.4
Proportion of times H_0: $\sigma_1^2 = \sigma_2^2$ was rejected ($\alpha = .05$)

Sample Sizes	Distribution				
	Normal	Uniform	t (df = 5)	Gamma (shape = 1)	Gamma (shape = .1)
(10, 10)	.054	.010	.121	.225	.693
(10, 20)	.056	.0068	.140	.236	.671
(20, 20)	.050	.0044	.150	.264	.673

7.4 Tests for Comparing $t > 2$ Population Variances

In the previous section, we discussed a method for comparing variances from two normally distributed populations based on taking independent random samples from the populations. In many situations, we will need to compare more than two populations. For example, we may want to compare the variability in the levels of nutrients in feed supplements from five different suppliers or the variability in scores of the students using SAT preparatory materials from the three major publishers of those materials. Thus, we need to develop a statistical test that will allow us to compare $t > 2$ population variances. The Brown-Forsythe-Levene (BFL) test is fairly complex in its computations, but it can be obtained from many of the statistical software packages. For example, R, SAS, and Minitab use the BFL test for comparing population variances.

The BFL test involves replacing the jth observation from sample i, y_{ij} with the random variable $z_{ij} = |y_{ij} - \tilde{y}_i|$, where $\tilde{y}_i$ is the sample median from the ith sample. The mean of all z_{ij}s is denoted $\bar{z}_{..}$, and the mean of the z_{ij}s from the ith sample is denoted $\bar{z}_{i.}$. With this notation, the BFL test statistic is computed as given in the following formula.

The BFL Test for Homogeneity of Population Variances

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$ homogeneity of variances

H_a: Population variances are not all equal

T.S.: $$L = \frac{\sum_{i=1}^{t} n_i(\bar{z}_{i.} - \bar{z}_{..})^2/(t-1)}{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_{i.})^2/(N-t)}$$

R.R.: For a specified value of α, reject H_0 if $L \geq F_{\alpha, df_1, df_2}$, where $df_1 = t - 1$, $df_2 = N - t$, $N = \sum_{i=1}^{t} n_i$, and F_{α, df_1, df_2} is the upper α percentile from the F distribution (from Table 8 in the Appendix).

Check assumptions and draw conclusions.

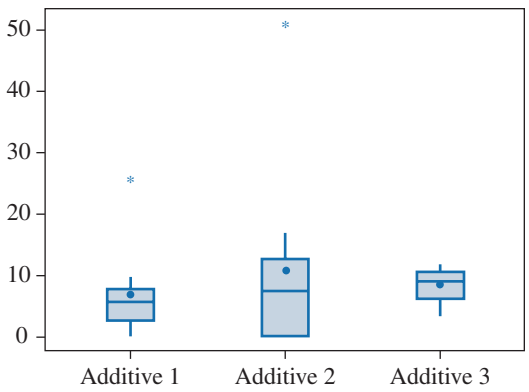
We will illustrate the computations for the BFL test in the following example. However, in most cases, we would recommend using a computer software package such as SAS or Minitab or R for conducting the test.

EXAMPLE 7.8

Three different additives that are marketed for increasing the miles per gallon (mpg) for automobiles were evaluated by a consumer testing agency. Past studies have shown an average increase of 8% in mpg for economy automobiles after using the product for 250 miles. The testing agency wanted to evaluate the variability in the increase in mileage over a variety of brands of cars within the economy class. The agency randomly selected 30 economy cars of similar age, number of miles on the odometer, and overall condition of the power train to be used in the study. It then randomly assigned 10 cars to each additive. The percentage increase in mpg obtained by each car was recorded for a 250-mile test drive. The testing agency wanted to evaluate whether there was a difference between the three additives with respect to their variability in the increase in mpg. The data are given here along with the intermediate calculations needed to compute the BFL's test statistic.

Solution Using the plots in Figures 7.11(a)–(d), we can observe that the samples from additive 1 and additive 2 do not appear to be samples from normally distributed populations. Hence, we should not use an F test for evaluating differences in the variances in this example. The information in Table 7.5 will assist us in calculating the value of the BFL test statistic. The medians of the percentage increase in mileage, y_{ij}s, for the three additives are 5.80, 7.55, and 9.15. We then calculate the absolute deviations of the data values about their respective medians—namely, $z_{1j} = |y_{1j} - 5.80|$, $z_{2j} = |y_{2j} - 7.55|$, and $z_{3j} = |y_{3j} - 9.15|$ for $j = 1, \ldots, 10$. These values are given in column 4 of the table. Next, we calculate the three means of these values, $\bar{z}_{1.} = 4.07$, $\bar{z}_{2.} = 8.88$, and $\bar{z}_{3.} = 2.23$. Next, we calculate the squared deviations of the z_{ij}s about their respective means, $(z_{ij} - \bar{z}_{i})^2$; that is, $(z_{1j} - 4.07)^2$, $(z_{2j} - 8.88)^2$, and $(z_{3j} - 2.23)^2$. These values are contained in column 6 of the table. Then we calculate the squared deviations of the z_{ij}s about the overall mean, $\bar{z}_{..} = 5.06$; that is, $(z_{ij} - \bar{z}_{i.})^2 = (z_{ij} - 5.06)^2$. The last column in the table contains these values. The final step is to sum columns 6 and 7, yielding

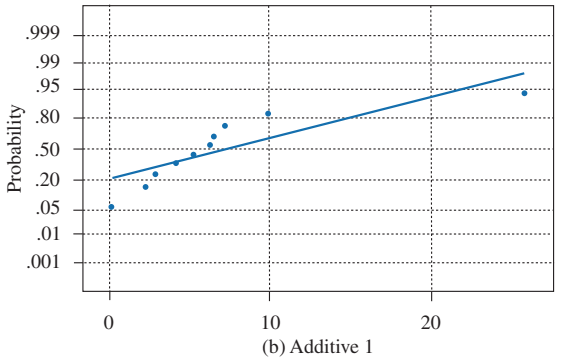
$$T_1 = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_{i.})^2 = 1{,}742.6 \quad \text{and} \quad T_2 = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_{..})^2 = 1{,}978.4$$



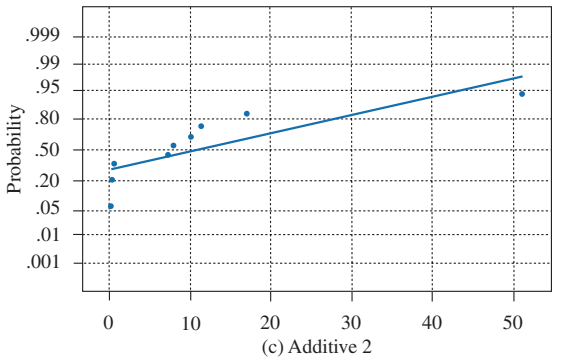
(a) Boxplots for three additives

FIGURE 7.11(a) Boxplots of additive 1, additive 2, and additive 3 (means are indicated by solid circles)

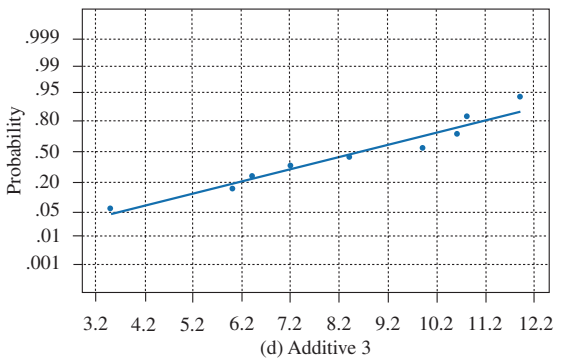
FIGURE 7.11(b)–(d)
Normal probability plots for additives 1, 2, and 3



(b) Additive 1



(c) Additive 2



(d) Additive 3

The value of BFL's test statistic, in an alternative form, is given by

$$L = \frac{(T_2 - T_1)/(t-1)}{T_1/(N-t)} = \frac{(1{,}978.4 - 1{,}742.6)/(3-1)}{1{,}742.6/(30-3)} = 1.827$$

The rejection region for the BFL test is this: Reject H_0 if $L \geq F_{\alpha,t-1,N-t} = F_{.05,2,27} = 3.35$. Because $L = 1.827$, we fail to reject H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ Using the R function ***pf***$(y, \text{df}_1, \text{df}_2) = P(F \leq y)$, the p-value is computed to be p-value $= P(F \geq 1.827) = 1 - $ ***pf***$(1.827, 2, 27) = .1802$, which is considerably larger than .05. Thus, there is insufficient evidence in the data to support the research hypothesis that there is a difference in the population variances of the percentage increases in mpg for the three additives.

TABLE 7.5
Percentage increase in mpg from cars driven using three additives

Additive	y_{1j}	$\tilde{y}_1$	$z_{1j} = \lvert y_{1j} - 5.80 \rvert$	$\bar{z}_{1.}$	$(z_{1j} - 4.07)^2$	$(z_{1j} - 5.06)^2$
1	4.2	5.80	1.60	4.07	6.1009	11.9716
1	2.9		2.90		1.3689	4.6656
1	0.2		5.60		2.3409	0.2916
1	25.7		19.90		250.5889	220.2256
1	6.3		0.50		12.7449	20.7936
1	7.2		1.40		7.1289	13.3956
1	2.3		3.50		0.3249	2.4336
1	9.9		4.10		0.0009	0.9216
1	5.3		0.50		12.7449	20.7936
1	6.5		0.70		11.3569	19.0096
Additive	y_{2j}	$\tilde{y}_2$	$z_{2j} = \lvert y_{2j} - 7.55 \rvert$	$\bar{z}_{2.}$	$(z_{2j} - 8.88)^2$	$(z_{2j} - 5.06)^2$
2	0.2	7.55	7.35	8.88	2.3409	5.2441
2	11.3		3.75		26.3169	1.7161
2	0.3		7.25		2.6569	4.7961
2	17.1		9.55		0.4489	20.1601
2	51.0		43.45		1,195.0849	1,473.7921
2	10.1		2.55		40.0689	6.3001
2	0.3		7.25		2.6569	4.7961
2	0.6		6.95		3.7249	3.5721
2	7.9		0.35		72.7609	22.1841
2	7.2		0.35		72.7609	22.1841
Additive	y_{3j}	$\tilde{y}_3$	$z_{3j} = \lvert y_{3j} - 9.15 \rvert$	$\bar{z}_{3.}$	$(z_{3j} - 2.23)^2$	$(z_{3j} - 5.06)^2$
3	7.2	9.15	1.95	2.23	0.0784	9.6721
3	6.4		2.75		0.2704	5.3361
3	9.9		0.75		2.1904	18.5761
3	3.5		5.65		11.6964	0.3481
3	10.6		1.45		0.6084	13.0321
3	10.8		1.65		0.3364	11.6281
3	10.6		1.45		0.6084	13.0321
3	8.4		0.75		2.1904	18.5761
3	6.0		3.15		0.8464	3.6481
3	11.9		2.75		0.2704	5.3361
				$\bar{z}_{..}$		
Total				5.06	1,742.6	1,978.4

7.5 RESEARCH STUDY: Evaluation of Methods for Detecting *E. coli*

A formal comparison between a new microbial method for the detection of *E. coli,* the Petrifilm HEC test, and an elaborate laboratory-based procedure, hydrophobic grid membrane filtration (HGMF), will now be described. The HEC test is easier to inoculate, more compact to incubate, and safer to handle than conventional procedures. However, it was necessary to compare the performance of the HEC test to that of the HGMF procedure in order to determine if the HEC test might be a viable method for detecting *E. coli.*

Defining the Problem

The developers of the HEC method sought answers to the following questions:

1. What parameters associated with the HEC and HGMF readings needed to be compared?
2. How many observations are necessary for a valid comparison of HEC and HGMF?
3. What type of experimental design would produce the most efficient comparison of HEC and HGMF?
4. What are the valid statistical procedures for making the comparisons?
5. What types of information should be included in a final report to document the evaluation of HEC and HGMF?

Collecting the Data

The experiment was designed to have two phases. Phase One of the study was to apply both procedures to pure cultures of *E. coli* representing 10^7 CFU/ml of strain E318N. Bacterial counts from both procedures would be obtained from a specified number of pure cultures. In order to determine the number of requisite cultures, the researchers decided on the following specification: The sample size would need to be large enough that there would be 95% confidence that the sample mean of the transformed bacterial counts would be within .1 units of the true mean for the HGMF transformed counts. From past experience with the HGMF procedure, the standard deviation of the transformed bacterial counts is approximately .25 units. The specification was made in terms of HGMF because there was no prior information concerning the counts from the HEC procedure. The following calculations yield the number of cultures needed to meet the specification.

The necessary sample size is given by

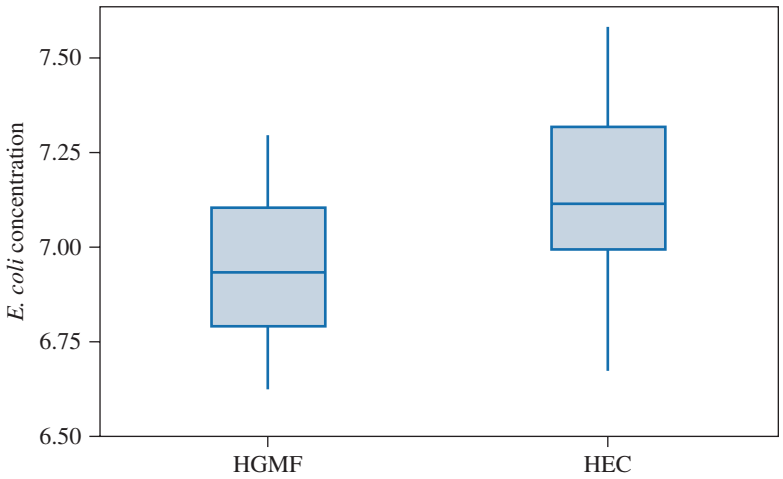
$$n = \frac{z^2_{\alpha/2}\hat{\sigma}^2}{E^2} = \frac{(1.96)^2(.25)^2}{(.1)^2} = 24.01$$

Based on the specified degree of precision in estimating the *E. coli* level, it was determined that the HEC and HGMF procedures would be applied to 24 pure cultures each. Thus, we have two independent samples of size 24 each. The determinations yielded the *E. coli* concentrations in transformed metric units ($\log_{10}$ CFU/ml) given in Table 7.6. (The values in Table 7.6 were simulated using the summary statistics given in the paper.)

TABLE 7.6
E. coli readings (Log_{10}CFU/ml) from HGMF and HEC

Sample	HGMF	HEC	Sample	HGMF	HEC
1	6.65	6.67	13	6.94	7.11
2	6.62	6.75	14	7.03	7.14
3	6.68	6.83	15	7.05	7.14
4	6.71	6.87	16	7.06	7.23
5	6.77	6.95	17	7.07	7.25
6	6.79	6.98	18	7.09	7.28
7	6.79	7.03	19	7.11	7.34
8	6.81	7.05	20	7.12	7.37
9	6.89	7.08	21	7.16	7.39
10	6.90	7.09	22	7.28	7.45
11	6.92	7.09	23	7.29	7.58
12	6.93	7.11	24	7.30	7.54

FIGURE 7.12
Boxplots of HEC and HGMF



The researchers would next prepare the data for a statistical analysis following the steps described in Section 2.5 of the textbook.

Summarizing the Data

The researchers were interested in determining if the two procedures yielded equivalent measures of *E. coli* concentrations. The boxplots of the experimental data are given in Figure 7.12. The two procedures appear to be very similar with respect to the width of box and length of whiskers, but HEC has a larger median than HGMF. The sample summary statistics are given here.

```
Descriptive Statistics: HGMF, HEC

Variable  N  N*    Mean    SE Mean   StDev
HGMF     24   0  6.9567     0.0414  0.2029
HEC      24   0  7.1383     0.0481  0.2358

Variable  Minimum    Q1    Median    Q3   Maximum
HGMF       6.6200  6.7900  6.9350  7.1050  7.3000
HEC        6.6700  6.9925  7.1100  7.3250  7.5800
```

From the summary statistics, we note that HEC yields a larger mean concentration than does HGMF. Also, the variability in concentration readings for HEC is greater than the value for HGMF. Our initial conclusion would be that the two procedures are yielding different distributions of readings for their determinations of *E. coli* concentration. However, we need to determine if the differences in their sample means and standard deviations imply a difference in the corresponding population values. We will next apply the appropriate statistical procedures in order to reach conclusions about the population parameters.

Analyzing the Data

Because the objective of the study was to evaluate the HEC procedure for its performance in detecting *E. coli,* it is necessary to evaluate its repeatability and its agreement with an accepted method for *E. coli*—namely, the HGMF procedure. Thus, we need to compare both the level and the variability in the two methods for determining *E. coli* concentrations. That is, we will need to test hypotheses about both the means and the standard deviations of HEC and HGMF *E. coli* concentrations. Recall we had 24 independent observations from the HEC and HGMF procedures on pure cultures of *E. coli* having a specified

level of 7 $\log_{10}$ CFU/ml. Prior to constructing confidence intervals or testing hypotheses, we must check whether the data represent random samples from normally distributed populations. From the boxplots displayed in Figure 7.12 and the normal probability plots in Figures 7.13(a)–(b), the data from both procedures appear to follow a normal distribution.

We next will test the hypotheses

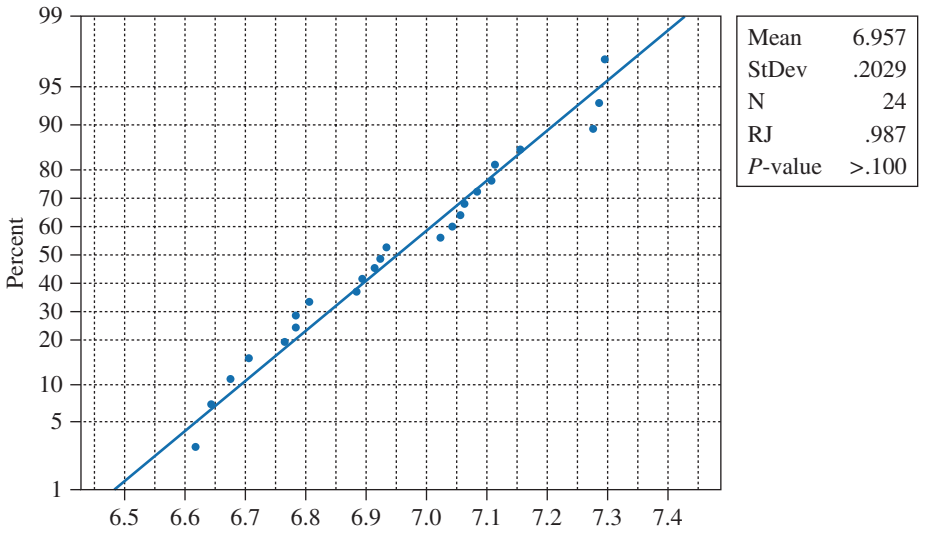
$$H_0: \sigma_1^2 = \sigma_2^2 \text{ versus } H_a: \sigma_1^2 \neq \sigma_2^2$$

where we designate HEC as population 1 and HGMF as population 2. The summary statistics are given in Table 7.7.

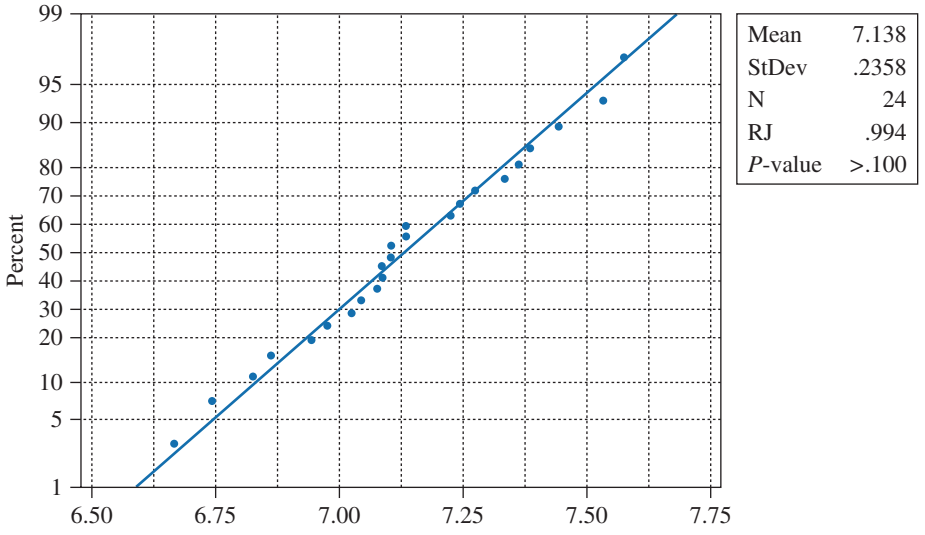
TABLE 7.7
HEC and HGMF summary statistics

Procedure	Sample Size	Mean	Standard Deviation
HEC	24	7.1383	.2358
HGMF	24	6.9567	.2029

FIGURE 7.13
Normal probability plots for HGMF and HEC



(a) *E. coli* concentration with HGMF



(b) *E. coli* concentration with HEC

R.R.: For a two-tailed test with $\alpha = .05$, we will reject H_0 if

$$F = \frac{s_1^2}{s_2^2} \leq F_{.975,23,23} = \frac{1}{F_{.025,23,23}} = \frac{1}{2.31} = .43 \quad \text{or} \quad F \geq F_{.025,23,23} = 2.31$$

Since $F = (.2358)^2/(.2029)^2 = 1.35$ is neither less than .43 nor greater than 2.31, we fail to reject H_0. The p-value is computed as follows: p-value $= \boldsymbol{pf}(\frac{1}{1.35}, 23, 23) + 1 - \boldsymbol{pf}(1.35, 23, 23) = .477$. Thus, we can conclude that HEC appears to have a degree of variability similar to that of HGMF in its determination of *E. coli* concentration. To obtain estimates of the variability in the HEC and HGMF readings, 95% confidence intervals on their standard deviations are given by

$$\left(\sqrt{\frac{(24-1)(.2358)^2}{38.08}}, \sqrt{\frac{(24-1)(.2358)^2}{11.69}}\right) = (0.18, .33) \text{ for } \sigma_{\text{HEC}}$$

and

$$\left(\sqrt{\frac{(24-1)(.2029)^2}{38.08}}, \sqrt{\frac{(24-1)(.2029)^2}{11.69}}\right) = (.16, .28) \text{ for } \sigma_{\text{HGMF}}$$

Because both the HEC and the HGMF *E. coli* concentration readings appear to be independent random samples from normal populations with a common standard deviation, we can use a pooled t test to evaluate

H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 \neq \mu_2$

R.R.: For a two-tailed test with $\alpha = .05$, we will reject H_0 if

$$|t| = \frac{|\bar{y}_1 - \bar{y}_2|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq t_{.025,46} = 2.01$$

Because $t = (7.14 - 6.96)/(.22\sqrt{\frac{1}{24} + \frac{1}{24}}) = 2.86$ is greater than 2.01, we reject H_0. The p-value = .006. Thus, there is significant evidence that the average HEC *E. coli* concentration readings differ from the average HGMF readings, with an estimated difference given by a 95% confidence interval on $\mu_{\text{HEC}} - \mu_{\text{HGMF}}$, (.05, .31). To estimate the average readings, 95% confidence intervals are given by (7.04,7.23) for μ_{HEC} and (6.86,7.04) for μ_{HGMF}. The HEC readings are on the average somewhat higher than the HGMF readings.

These findings would then prepare us for the second phase of the study. In this phase, HEC and HGMF will be applied to the same sample of meats in a research study similar to what would be encountered in a meat-monitoring setting. The two procedures had similar levels of variability, but HEC produced *E. coli* concentration readings higher than those of HGMF. Thus, the goal of Phase Two would be to calibrate the HEC readings to the HGMF readings. We will discuss this phase of the study in Chapter 11.

Reporting the Conclusions

We would need to write a report summarizing our findings concerning Phase One of the study. We would need to include the following:

1. Statement of objective for study
2. Description of study design and data collection procedures
3. Numerical and graphical summaries of data sets
4. Description of all inference methodologies
   - t and F tests
   - t-based confidence intervals on means

- chi-square–based confidence intervals on standard deviations
- verification that all necessary conditions for using inference techniques were satisfied

5. Discussion of results and conclusions
6. Interpretation of findings relative to previous studies
7. Recommendations for future studies
8. Listing of data sets

7.6 Summary and Key Formulas

In this chapter, we discussed procedures for making inferences concerning population variances or, equivalently, population standard deviations. Estimation and statistical tests concerning σ make use of the chi-square distribution with $df = n - 1$. Inferences concerning the ratio of two population variances or standard deviations utilize the F distribution with $df_1 = n_1 - 1$ and $df_2 = n_1 - 2$. Finally, when we developed tests concerning differences in $t > 2$ population variances, we used the Brown-Forsythe-Levene (BFL) test statistic.

The need for inferences concerning one or more population variances can be traced to our discussion of numerical descriptive measures of a population in Chapter 3. To describe or make inferences about a population of measurements, we cannot always rely on the mean, a measure of central tendency. Many times in evaluating or comparing the performance of individuals on a psychological test, the consistency of manufactured products emerging from a production line, or the yields of a particular variety of corn, we gain important information by studying the population variance.

Key Formulas

1. $100(1 - \alpha)\%$ confidence interval for σ^2 (or σ)

$$\left(\frac{(n-1)s^2}{\chi^2_U}, \frac{(n-1)s^2}{\chi^2_L}\right)$$

or

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_U}}, \sqrt{\frac{(n-1)s^2}{\chi^2_L}}\right)$$

2. Statistical test for σ^2 (σ_0^2 specified)

$$\text{T.S.: } \chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

3. Statistical test for σ_1^2/σ_2^2

$$\text{T.S.: } F = \frac{s_1^2}{s_2^2}$$

4. $100(1 - \alpha)\%$ confidence interval for σ_1^2/σ_2^2 (or σ_1/σ_2)

$$\left(\frac{s_1^2}{s_2^2}F_L, \frac{s_1^2}{s_2^2}F_U\right)$$

where

$$F_L = \frac{1}{F_{\alpha/2, df_1, df_2}} \quad \text{and} \quad F_U = F_{\alpha/2, df_2, df_1}$$

or

$$\left(\sqrt{\frac{s_1^2}{s_2^2}F_L}, \sqrt{\frac{s_1^2}{s_2^2}F_U}\right)$$

5. Statistical test for

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$$

BFL test should be used.

$$\text{T.S.: } L = \frac{\sum_{i=1}^{t} n_i(\bar{z}_i - \bar{z}_{..})^2/(t-1)}{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_{i.})^2/(N-t)}$$

where $z_{ij} = |y_{ij} - \tilde{y}_{i.}|$, $\tilde{y}_{i.}$ = median $(y_{i1}, \ldots, y_{in_i})$, $\bar{z}_{i.}$ = mean $(z_{i1}, \ldots, z_{in_i})$, and $\bar{z}_{..}$ = mean $(z_{11}, \ldots, z_{tn_t})$

7.7 Exercises

7.1 Introduction

Env. **7.1** For the *E. coli* research study, answer the following.

- **a.** What are the populations of interest?
- **b.** What are some factors other than the type of detection method (HEC versus HGMF) that may cause variation in the *E. coli* readings?
- **c.** Describe a method for randomly assigning the *E. coli* samples to the two devices for analysis.
- **d.** State several hypotheses that may be of interest to the researchers.

7.2 Estimation and Tests for a Population Variance

Basic **7.2** Suppose a random variable W has a chi-square distribution with df = 23. Determine the following probabilities.

- **a.** $P(W > 41.64)$
- **b.** $P(W > 35.17)$
- **c.** $P(W \leq 13.09)$
- **d.** $P(W \leq 12.14)$
- **e.** $P(W \leq 35.17)$
- **f.** $P(12.14 < W \leq 35.17)$

Basic **7.3** Find the following percentiles for a chi-square distribution with df = 18.

- **a.** $\chi^2_{.05}$
- **b.** $\chi^2_{.01}$
- **c.** $\chi^2_{.95}$
- **d.** $\chi^2_{.025}$
- **e.** $\chi^2_{.03}$
- **f.** $\chi^2_{.94}$

Basic **7.4** Table 7 in the Appendix is useful for obtaining percentiles for the chi-square distribution for a wide range of values of degrees of freedom and values of α. Alternatively, when a computer is available, a software program such as R can be used to obtain percentiles or to compute p-values for values of degrees of freedom and values of α not provided in Table 7. However, in those situations when a computer is unavailable or Table 7 does not have the desired percentile for a specified degrees of freedom, the following approximation can be used provided df > 40.

$$\chi^2_\alpha \approx \nu\left(1 - \frac{2}{9\nu} + z_\alpha\sqrt{\frac{2}{9\nu}}\right)^3$$

where χ^2_α is the upper percentile of the chi-square distribution with df $= \nu$ and z_α is the upper α percentile of a standard normal distribution.

- **a.** For df = 20, compare the values obtained from the approximation for $\chi^2_{.05}$ and $\chi^2_{.95}$ to the values listed in Table 7.
- **b.** For df = 60, compare the values obtained from the approximation for $\chi^2_{.05}$ and $\chi^2_{.95}$ to the values listed in Table 7.
- **c.** For df = 90, compare the values obtained from the approximation for $\chi^2_{.05}$ and $\chi^2_{.95}$ to the values listed in Table 7.
- **d.** For df = 240, compare the values obtained from the approximation for $\chi^2_{.05}$ and $\chi^2_{.95}$ to the values listed in Table 7.
- **e.** Comment on the accuracy of the approximation for the percentiles obtained in parts (a)–(d).

Bus. **7.5** A production process for filling orange juice containers labeled as 64 ounces is monitored for the actual amount of juice in the container. The process is designed such that the amount of juice in the containers has a normal distribution with a mean of 64.3 ounces and a standard deviation of .15 ounces. The process is monitored by randomly selecting 24 containers every hour and measuring the actual amount of juice in the containers. An increase in the standard

deviation beyond .15 ounces with the mean remaining at 64.3 ounces will result in a production run with too many underfilled and overfilled containers. The following data are the actual amounts of juice in a random sample of 24 containers.

64.37	64.26	64.22	64.42	64.13	64.44	64.64	64.19	63.85	64.17	64.21	64.23
64.64	64.12	63.98	64.34	64.20	64.31	64.15	64.09	64.33	64.19	64.57	64.19

a. If the amount of juice in the containers has a normal distribution with a mean of 64.3 ounces and a standard deviation of .15 ounces, what proportion of containers filled on the production line will be underfilled (contain less than 64 ounces)? What percentage will be overfilled?
b. Construct a 95% confidence interval on the process standard deviation.
c. Do the data indicate that the process standard deviation is greater than .15 ounces? Use $\alpha = .05$ in reaching your conclusion.
d. What is the p-value of your test?
e. Is there any indication that the necessary conditions for constructing the confidence interval and conducting the test may be violated?
f. What is the population about which inferences can be made using the given data?

Engin. **7.6** A leading researcher in the study of interstate highway accidents proposes that a major cause of many collisions on the interstates is not the speed of the vehicles but rather the *difference* in speeds of the vehicles. When some vehicles are traveling slowly while other vehicles are traveling at speeds greatly in excess of the speed limit, the faster-moving vehicles may have to change lanes quickly, which can increase the chance of an accident. Thus, when there is a large variation in the speeds of the vehicles in a given location on the interstate, there may be a larger number of accidents than when the traffic is moving at a more uniform speed. The researcher believes that when the standard deviation in speed of vehicles exceeds 10 mph, the rate of accidents is greatly increased. During a 1-hour period of time, a random sample of 50 vehicles is selected from a section of an interstate known to have a high rate of accidents, and their speeds are recorded using a radar gun. The data are presented here.

56.1	57.0	53.9	50.2	54.2	47.9	78.1	60.2	47.4	68.8
45.5	63.3	59.7	74.3	61.4	58.7	61.2	64.7	64.3	48.2
57.7	72.1	72.0	67.6	47.6	65.9	72.3	55.7	55.0	75.2
62.8	47.0	48.1	62.9	64.0	80.6	51.2	53.7	53.3	58.3
68.2	69.5	51.8	68.8	63.8	61.8	59.3	63.6	54.7	59.9

a. Do the data indicate any violations in the conditions necessary to use the chi-square procedures for generating confidence intervals and testing hypotheses?
b. Estimate the standard deviation in the speeds of the vehicles traveling on the interstate using a 95% confidence interval.
c. Do the data indicate that the standard deviation in vehicle speeds exceeds 10 mph? Use $\alpha = .05$ in reaching your conclusion.
d. To what population can the inferences obtained in parts (b) and (c) be validly applied?

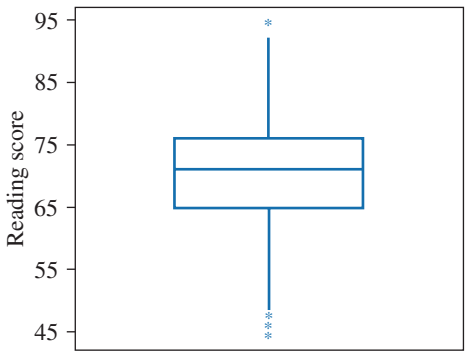
Edu. **7.7** A large public school system was evaluating its elementary school reading program. In particular, educators were interested in the performance of students on a standardized reading test given to all third graders in the state. The mean score on the test was compared to the state average to determine the school system's rating. Also, the educators were concerned with the variation in scores. If the mean scores were at an acceptable level but the variation was high, this would indicate that a large proportion of the students still needed remedial reading programs. Also, a large variation in scores might indicate a need for programs for those students at the gifted level. Without accelerated reading programs, these students lose interest during reading classes. To obtain information about students early in the school year (the statewide test is given during the last month of the school year), a random sample of 150 third-grade students was given the exam used in the previous year. The possible scores on the reading test range from 0 to 100. The data are summarized here.

```
Descriptive Statistics for Reading Scores

Variable         N      Mean   Median   TrMean   StDev   SE Mean
Reading        150    70.571   71.226   70.514   9.537     0.779

Variable   Minimum   Maximum       Q1       Q3
Reading     44.509    94.570   65.085   76.144
```

a. Does the following plot of the data suggest any violation of the conditions necessary to use the chi- square procedures for generating a confidence interval and a test of hypotheses about σ?

b. Estimate the variation in reading scores using a 99% confidence interval.

c. Do the data indicate that the standard deviation in reading scores is greater than 9, the standard deviation for all students taking the exam the previous year? Use $\alpha = .01$ in reaching your conclusion.



Edu. **7.8** Refer to Exercise 7.7.

a. Compute the p-value of the test conducted in Exercise 7.7.

b. If the value of α is increased to .05, would the conclusion reached in Exercise 7.7 change?

Engin. **7.9** Baseballs vary somewhat in their rebounding coefficient. A baseball that has a large rebound coefficient will travel farther when the same force is applied to it than a ball with a smaller coefficient. To achieve a game in which each batter has an equal opportunity to hit a home run, the balls should have nearly the same rebound coefficient. A standard test has been developed to measure the rebound coefficient of baseballs. A purchaser of large quantities of baseballs requires that the mean coefficient value be 85 units and the standard deviation be less than 2 units. A random sample of 40 baseballs is selected from a large batch of balls and tested. The data are given here.

84.8	88.1	85.1	88.0	86.6	85.3	85.1	91.4	83.4	87.2
83.7	89.5	85.6	83.5	81.6	81.1	83.6	81.2	84.7	87.0
87.5	84.3	86.9	83.3	85.9	82.2	88.2	83.5	82.7	86.0
87.3	87.9	82.6	80.5	85.6	82.3	79.3	84.9	80.6	83.9

a. Do the data indicate any violations in the conditions necessary to use the chi-square procedures for generating confidence intervals and testing hypotheses?

b. Estimate the standard deviation in the rebound coefficients using a 99% confidence interval.

c. Do the data indicate that the mean rebound coefficient is less than 85? Use $\alpha = .05$ in reaching your conclusion.

d. Do the data indicate that the standard deviation in rebound coefficients exceeds 2? Use $\alpha = .01$ in reaching your conclusion.

e. To what population can the inferences obtained in parts (b)–(d) be validly applied?

7.10 Use the results of the simulation study, summarized in Table 7.2, to answer the following questions.

a. Which of skewness or heavy-tailedness appears to have the stronger effect on the chi-square tests?

b. For a given population distribution, does increasing the sample size yield α values more nearly equal to the nominal value of .05? Justify your answer, and provide reasons why this may occur.

c. For the short-tailed distribution (uniform), the actual probability of Type I error is smaller than the specified value of .05. Provide both a negative and a positive impact on the chi-square test of having a decrease in the specified value of α.

7.3 Estimation and Tests for Comparing Two Population Variances

Basic **7.11** Find the value that locates an area α in the upper tail of the F distribution; that is, find $F\alpha$ for the following values of α and degrees of freedom.

a. $\alpha = .05$, $df_1 = 7$, $df_2 = 9$

b. $\alpha = .025$, $df_1 = 9$, $df_2 = 7$

c. $\alpha = .01$, $df_1 = 17$, $df_2 = 9$

d. $\alpha = .10$, $df_1 = 9$, $df_2 = 20$

e. $\alpha = .25$, $df_1 = 15$, $df_2 = 12$

f. $\alpha = .15$, $df_1 = 15$, $df_2 = 19$

Basic **7.12** Find the value that locates an area α in the upper tail of the F distribution; that is, find F_α for the following values of α and degrees of freedom.

a. $\alpha = .05$, $df_1 = 6$, $df_2 = 45$

b. $\alpha = .025$, $df_1 = 8$, $df_2 = 55$

c. $\alpha = .01$, $df_1 = 7$, $df_2 = 38$

d. $\alpha = .10$, $df_1 = 12$, $df_2 = 87$

e. $\alpha = .005$, $df_1 = 7$, $df_2 = 46$

f. $\alpha = .001$, $df_1 = 15$, $df_2 = 58$

Basic **7.13** Find the following percentiles for an F distribution with the following specifications:

a. $\alpha = .05$, $df_1 = 14$, $df_2 = 9$

b. $\alpha = .025$, $df_1 = 39$, $df_2 = 27$

c. $\alpha = .01$, $df_1 = 50$, $df_2 = 39$

d. $\alpha = .10$, $df_1 = 39$, $df_2 = 40$

e. $\alpha = .001$, $df_1 = 45$, $df_2 = 45$

f. $\alpha = .005$, $df_1 = 25$, $df_2 = 39$

Basic **7.14** Random samples of sizes $n_1 = 25$ and $n_2 = 20$ were selected from populations A and B, respectively. From the samples, the standard deviations were computed to be $s_1 = 5.2$ and $s_2 = 6.8$.

a. Do the data provide substantial evidence to indicate the populations have different standard deviations? Use $\alpha = .05$.

b. Estimate the relative sizes of the standard deviations by constructing a 95% confidence interval for the ratio of the standard deviations σ_1/σ_2.

c. The data and populations must satisfy what conditions in order for your test and confidence interval to be valid?

Engin. **7.15** A soft-drink firm is evaluating an investment in a new type of canning machine. The company has already determined that it will be able to fill more cans per day for the same cost if the new machines are installed. However, it must determine the variability of fills using the new machines and wants the variability from the new machines to be equal to or smaller than that currently obtained using the old machines. A study is designed in which random samples of 40 cans are selected from the output of both types of machines and the amount of fill (in ounces) is determined. The data are given below.

Old Machine								New Machine							
16.74	15.75	16.19	16.54	15.92	16.29	16.44	16.29	15.64	15.81	16.20	16.36	16.36	16.05	16.07	16.04
16.38	16.47	16.56	16.42	16.08	16.47	16.02	16.74	16.08	16.31	16.50	16.14	16.12	16.30	16.41	16.11
15.97	16.47	16.06	16.64	16.40	16.40	16.28	16.66	16.20	16.29	15.75	16.22	16.12	16.23	16.19	16.59
16.80	16.36	16.36	16.27	16.43	16.26	16.31	16.59	16.08	16.07	16.15	16.50.	16.25	16.25	16.19	16.13
16.24	16.63	16.15	16.17	16.32	16.81	16.27	17.09	15.96	16.02	16.29	15.99	15.99	16.42	16.15	16.23

a. Estimate the standard deviations in fill for types of machines using 95% confidence intervals.

b. Do these data present sufficient evidence to indicate that the new type of machine has less variability of fills than the old machine?

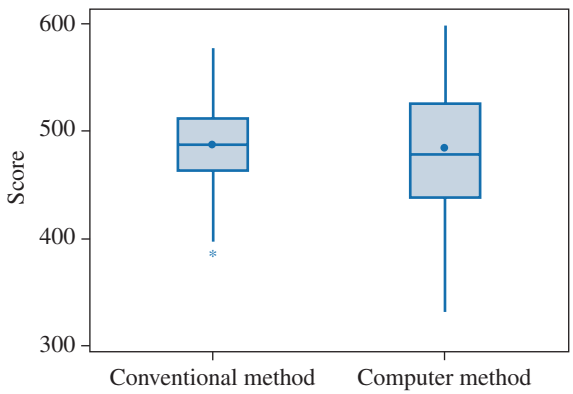
c. Do the necessary conditions for conducting the inference procedures in parts (a) and (b) appear to be satisfied? Justify your answer.

Edu.

7.16 The SAT Reasoning Test is an exam taken by most high school students as part of their college admission requirements. A proposal has been made to alter the exam by having the students take the exam on a computer. The exam questions would be selected for the student in the following fashion. For a given section of questions, if the student answers the initial questions posed correctly, then the following questions become increasingly difficult. If the student provides incorrect answers for the initial questions asked in a given section, then the level of difficulty of latter questions does not increase. The final score on the exams will be standardized to take into account the overall difficulty of the questions on each exam. The testing agency wants to compare the scores obtained using the new method of administering the exam to the scores using the current method. A group of 182 high school students is randomly selected to participate in the study with 91 students randomly assigned to each of the two methods of administering the exam. The data are summarized in the following table and boxplots for the math portion of the exam.

	Summary Data for SAT Reasoning Exams		
Testing Method	**Sample Size**	**Mean**	**Standard Deviation**
Computer	91	484.45	53.77
Conventional	91	487.38	36.94

Boxplots of conventional and computer methods (means are indicated by solid circles)



Evaluate the two methods of administering the SAT exam. Provide tests of hypotheses and confidence intervals. Are the means and standard deviations of scores for the two methods equivalent? Justify your answer using $\alpha = .05$.

7.17 Use the results of the simulation study summarized in Table 7.4 to answer the following questions.

a. Which of skewness or heavy-tailedness appears to have the stronger effect on the F tests?

b. For a given population distribution, does increasing the sample size yield α values more nearly equal to the nominal value of .05? Justify your answer, and provide reasons why this may occur.

c. For the short-tailed distribution (uniform), the actual probability of Type I error is smaller than the specified value of .05. Provide both a negative and a positive impact on the F test of having a decrease in the specified value of α.

7.4 Tests for Comparing $t > 2$ Population Variances

Bio. **7.18** A wildlife biologist was interested in determining the effect of raising deer in captivity on the size of the deer. She decided to consider three populations: deer raised in the wild, deer raised on large hunting ranches, and deer raised in zoos. She randomly selected eight deer in each of the three environments and weighed the deer at age 1 year. The weights (in pounds) are given in the following table.

Environment	Weight (in pounds) of Deer							
Wild	114.7	128.9	111.5	116.4	134.5	126.7	120.6	129.59
Ranch	120.4	91.0	119.6	119.4	150.0	169.7	100.9	76.1
Zoo	103.1	90.7	129.5	75.8	182.5	76.8	87.3	77.3

a. The biologist hypothesized that the weights of deer from captive environments would have a larger level of variability than the weights from deer raised in the wild. Do the data support her contention?
b. Are the requisite conditions for the test you used in part (a) satisfied in this situation? Provide plots to support your answer.

Theory **7.19** Why do you think that the BFL test is effective in testing for differences in the variances from populations having nonnormal distributions, whereas the F statistic cannot be applied to nonnormal distributions?

Supplementary Exercises

Bus. **7.20** A consumer-protection magazine was interested in comparing tires purchased from two different companies that each claimed their tires would last 40,000 miles. A random sample of 10 tires of each brand was obtained and tested under simulated road conditions. The number of miles until the tread thickness reached a specified depth was recorded for all tires. The data are given next (in thousands of miles).

Brand I	38.9	39.7	42.3	39.5	39.6	35.6	36.0	39.2	37.6	39.5
Brand II	44.6	46.9	48.7	41.5	37.5	33.1	43.4	36.5	32.5	42.0

a. Plot the data, and compare the distributions of longevity for the two brands.
b. Construct 95% confidence intervals on the means and standard deviations for the number of miles until tread wearout occurred for both brands.
c. Does there appear to be a difference in wear characteristics for the two brands? Justify your statement with appropriate plots of the data, tests of hypotheses, and confidence intervals.

Med. **7.21** A pharmaceutical company manufactures a particular brand of antihistamine tablets. In the quality control division, certain tests are routinely performed to determine whether the product being manufactured meets specific performance criteria prior to release of the product onto the market. In particular, the company requires that the potencies of the tablets lie in the range of 90% to 110% of the labeled drug amount.

a. If the company is manufacturing 25 mg tablets, within what limits must tablet potencies lie?
b. A random sample of 30 tablets is obtained from a recent batch of antihistamine tablets. The data for the potencies of the tablets are given next. Is the assumption of normality warranted for inferences about the population variance?
c. Translate the company's 90% to 110% specifications on the range of the product potency into a statistical test concerning the population variance for potencies. Draw conclusions based on $\alpha = .05$.

24.1	27.2	26.7	23.6	26.4	25.2
25.8	27.3	23.2	26.9	27.1	26.7
22.7	26.9	24.8	24.0	23.4	25.0
24.5	26.1	25.9	25.4	22.9	24.9
26.4	25.4	23.3	23.0	24.3	23.8

Bus. **7.22** The risk of an investment is measured in terms of the variance in the return that could be observed. Random samples of 10 yearly returns were obtained from two different portfolios. The data are given next (in thousands of dollars).

Portfolio 1	130	135	135	131	129	135	126	136	127	132
Portfolio 2	154	144	147	150	155	153	149	139	140	141

a. Does portfolio 2 appear to have a higher risk than portfolio 1?
b. Give a p-value for your test, and place a confidence interval on the ratio of the standard deviations of the two portfolios.
c. Provide a justification that the required conditions have been met for the inference procedures used in parts (a) and (b).

7.23 Refer to Exercise 7.22. Are there any differences in the average returns for the two portfolios? Indicate the method you used in arriving at a conclusion, and explain why you used it.

Med. **7.24** Sales from weight-reducing agents marketed in the United States represent sizable amounts of income for many of the companies that manufacture these products. Psychological as well as physical effects often contribute to how well a person responds to the recommended therapy. Consider a comparison of two weight-reducing agents, A and B. In particular, consider the length of time people remain on the therapy. A total of 26 overweight males, matched as closely as possible physically, were randomly divided into two groups. Those in group 1 received preparation A and those assigned to group 2 received preparation B. The data are given here (in days).

Preparation A	42	47	12	17	26	27	28	26	34	19	20	27	34
Preparation B	35	38	35	36	37	35	29	37	31	31	30	33	44

Compare the lengths of times that people remain on the two therapies. Make sure to include all relevant plots, tests, confidence intervals, and a written conclusion concerning the two therapies.

7.25 Refer to Exercise 7.24. How would your inference procedures change if preparation A was an old product that had been on the market a number of years and preparation B was a new product, and we wanted to determine whether people would continue to use B a longer time in comparison to preparation A?

Gov. **7.26** A school district in a midsized city currently has a single high school for all its students. The number of students attending the high school has become somewhat unmanageable, and, hence, the school board has decided to build a new high school. The school board after considerable deliberation divides the school district into two attendance zones, one for the current high school and one for the new high school. The board guaranteed the public that the mean family income was the same for the two zones. However, a group of parents is concerned that the two zones have greatly different family socioeconomic distributions. A random sample of 30 homeowners were selected from each zone to be interviewed concerning relevant family traits. Two families in zone II refused to participate in the study, even though the researcher promised to keep interview information confidential. One aspect of the collected data was family income. The incomes, in thousands of dollars, produced the following data.

Zone I Incomes									
44.1	69.0	46.9	41.7	61.3	43.9	48.0	61.3	31.2	49.3
57.1	46.5	53.6	47.0	47.0	53.7	39.2	64.3	40.9	45.4
58.2	54.6	66.6	36.6	58.2	45.8	62.9	53.2	56.1	53.0

Zone II Incomes									
53.6	58.4	56.1	48.1	56.5	50.2	60.0	44.4	56.5	57.3
58.6	53.1	60.4	54.2	54.2	59.5	54.3	59.2	53.9	48.8
58.1	58.4	51.7	59.3	51.4	56.3	57.7	54.3	62.1	47.9

a. Verify that the two attendance zones have the same mean income.
b. Use these data to test the hypothesis that although the mean family incomes are nearly the same in the two zones, zone I has a much higher level of variability than zone II in terms of family income.
c. Place a 95% confidence interval on the ratio of the two standard deviations.
d. For each zone, use your estimates of the zone standard deviations to determine the range of incomes that would contain 95% of all incomes in each of the zones.
e. Verify that the necessary conditions have been met to apply the procedures you used in parts (a)–(c).

Engin. **7.27** Refer to Example 6.2. In this example, the pooled t-based confidence interval procedures were used to estimate the difference between domestic and imported mean repair costs. Verify that the necessary conditions were satisfied.

Bus. **7.28** Refer to Exercise 6.59. The company officials decided to use the separate-variance t test in deciding whether the mean potency of the drug after 1 year of storage was different from the mean potency of the drug from current production. Provide evidence that their decision in fact was correct.

Engin. **7.29** A casting company has several ovens in which they heat the raw materials prior to pouring them into a wax mold. It is very important that these metals be heated to a precise temperature with very little variation. Three ovens are selected at random, and their temperatures are recorded (°C) very acurately on 10 successive heats. The collected data are as follows:

Oven	Temperature °C									
1	1,670.87	1,670.88	1,671.51	1,672.01	1,669.63	1,670.95	1,668.70	1,671.86	1,669.12	1,672.52
2	1,669.16	1,669.60	1,669.76	1,669.18	1,671.92	1,669.69	1,669.45	1,669.35	1,671.89	1,673.45
3	1,673.08	1,672.75	1,675.14	1,674.94	1,671.33	1,660.38	1,679.94	1,660.51	1,668.78	1,664.32

a. Is there significant evidence ($\alpha = .05$) that the three ovens have different levels of variation in their temperatures?
b. Assess the order of magnitude in the differences in standard deviations by placing 95% confidence intervals on the ratios of the three pairs of standard deviations.
c. Do the conditions that are required by your statistical procedures in parts (a) and (b) appear to be valid?

Med. **7.30** A new steroidal treatment for a skin condition in dogs was under evaluation by a veterinary hospital. One of the possible side effects of the treatment is that a dog receiving the treatment may have an allergic reaction to the treatment. This type of allegeric reaction manifests itself through an elevation in the resting pulse rate of the dog after the dog has received the treatment for a period of time. A group of 80 dogs of the same breed and age, and all having the skin condition, is randomly assigned to either a placebo treatment or the steroidal treatment. Four days after receiving the treatment, either steroidal or placebo, resting pulse rate measurements are taken on all the dogs. These data are displayed here. Dogs of this age and breed have a fairly constant resting pulse rate of 100 beats a minute. The researchers are interested in testing whether there is a significant difference between the placebo and treatment dogs in terms of both the means and standard deviations of the resting pulse rates.

Placebo Group Pulse Rates								
105.1	103.3	102.1	102.3	101.5	100.6	104.5	103.2	101.8
102.1	108.1	103.2	104.0	103.9	105.3	103.6	102.3	103.9
103.0	107.0	102.3	103.5	111.7	101.4	103.0	101.1	103.7
102.3	106.2	100.8	102.1	104.3	104.0	102.2	103.1	104.7
102.3	110.1	103.1	103.4					

Treatment Group Pulse Rates								
107.6	107.8	110.4	106.6	108.2	113.4	113.5	108.7	108.2
106.0	105.3	107.1	110.3	108.7	107.4	111.1	105.9	106.9
106.4	111.5	106.8	107.8	106.1	106.7	105.0	110.4	105.9
106.4	106.0	106.0	106.9	107.6	107.0	105.8	108.6	109.3
108.5	106.9	107.0	109.2					

a. Is there significant evidence of an increase in the mean pulse rates for those dogs receiving the treatment?
b. Is there significant evidence of a difference in the levels of variability in pulse rate between the placebo and the treatment group of dogs?
c. Provide a 95% confidence interval on the difference in mean pulse rates between the placebo and treatment groups.
d. Do the necessary conditions hold for the statistical procedures you applied in parts (a)–(c)? Justify your answer.

Met. **7.31** A series of experiments was designed to test a hypothesis that massive silver iodide seeding can, under specified conditions, lead to increased precipitation. The data from these experiments were reported in the article ***"A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification" [Technometrics (1975) 17:161–166]***. The rain volume falling from the cloud after seeding with silver iodide is reported here.

Rainfall (acre-feet) Unseeded Clouds							
129.6	31.4	2745.6	489.1	430.0	302.8	119.0	4.1
92.4	17.5	200.7	274.7	274.7	7.7	1656.0	978.0
198.6	703.4	1697.8	334.1	118.3	255.0	115.3	242.5
32.7	40.6						

Rainfall (acre-feet) Seeded Clouds							
26.1	26.3	87.0	95.0	372.4	0.0	17.3	24.4
11.5	321.2	68.5	81.2	47.3	28.6	830.1	345.5
1202.6	36.6	4.9	4.9	41.1	29.0	163.0	244.3
147.8	21.7						

a. Is there significant evidence that seeding has increased the mean rainfall?
b. Is there a significant difference in the levels of variability in the amount of rainfall between seeded and unseeded clouds?
c. In order for seeding to be economically viable, it must on the average produce at least 100 more acre-feet of rainfall over usual (unseeded) rainfall. Is there evidence in this data set that seeding is economically viable?

7.32 Refer to the epilepsy data in Table 3.19. The researchers were interested in determining whether the treatment patients and placebo patients had differences in the number of epileptic seizures during their fourth clinic visit after receiving either the treatment or a placebo.

a. Is there significant evidence that the mean number of seizures is smaller in the treatment group than in the placebo group?
b. Compare the treatment and placebo groups relative to the variation in their respective number of seizures during the fourth visit.
c. Do you think that the treatment was effective? Justify your answer.

Soc. **7.33** Refer to Exercise 3.55.

a. What are the target populations for this study?
b. The state agency in charge of allocations for food stamps wants to deterinine if the level of variation in expenditures differed for the five groups. Conduct a test and construct confidence intervals to answer the agency's question.

CHAPTER 8

Inferences About More Than Two Population Central Values

8.1 Introduction and Abstract of Research Study

In Chapter 6, we presented methods for comparing two population means based on independent random samples selected from each of the populations. In many practical/scientific settings, the number of populations for which we want to make comparisons will be three or more. For example, it is claimed that the influx of undocumented workers into the United States has resulted in the suppression of wages of laborers, especially in the southwestern states. Advocates for the unionization of farm workers argue that it is not the documentation status of the workers that is causing the decrease in wages but rather the lack of union representation. We wish to compare the mean hourly wage for farm laborers from three different classifications (union-documented, nonunion-documented, nonunion-undocumented). Independent random samples of farm laborers would be selected from each of the three classifications (populations). The sample means and sample variances would then be used to make an inference about the corresponding population mean hourly wages. It is almost certain that the sample means would differ; however, this does not necessarily imply a difference among the population

means. How do we determine the size of difference in the sample means necessary for us to state with some degree of certainty that the population means are different? The statistical procedure called *analysis of variance* will provide us with the answer to this question.

The reason we call the testing procedure an analysis of variance will be demonstrated by using the hourly wage example discussed in the previous paragraph. Assume that we wish to compare the mean hourly wages of the three classifications of farm laborers. We will use a random sample of five workers from each of the populations to illustrate the basic ideas of an analysis of variance. The sample size is unreasonably small for a real evaluation of wages, but it is used in order to simplify the presentation.

Suppose the sample data (hourly wages, in dollars) are as shown in Table 8.1. Do these data present sufficient evidence to indicate differences among the three population means? A brief visual inspection of the data indicates very little variation within a sample, whereas the variability among the sample means is much larger. Because the variability among the sample means is large *in comparison to the* **within-sample variation,** we might conclude intuitively that the corresponding population means are different.

within-sample variation

Table 8.2 illustrates a situation in which the sample means are the same as given in Table 8.1, but the variability within a sample is much larger, and the **between-sample variation** is small relative to the within-sample variability. We would be less likely to conclude that the corresponding population means differ based on these data.

between-sample variation

The variations in the two sets of data, Tables 8.1 and 8.2, are shown graphically in Figure 8.1. The strong evidence to indicate a difference in population

TABLE 8.1
A comparison of three sample means (small amount of within-sample variation)

Sample from Population		
1	**2**	**3**
5.90	5.51	5.01
5.92	5.50	5.00
5.91	5.50	4.99
5.89	5.49	4.98
5.88	5.50	5.02
$\bar{y}_1 = 5.90$	$\bar{y}_2 = 5.50$	$\bar{y}_3 = 5.00$

TABLE 8.2
A comparison of three sample means (large amount of within-sample variation)

Sample from Population		
1	**2**	**3**
5.90	6.31	4.52
4.42	3.54	6.93
7.51	4.73	4.48
7.89	7.20	5.55
3.78	5.72	3.52
$\bar{y}_1 = 5.90$	$\bar{y}_2 = 5.50$	$\bar{y}_3 = 5.00$

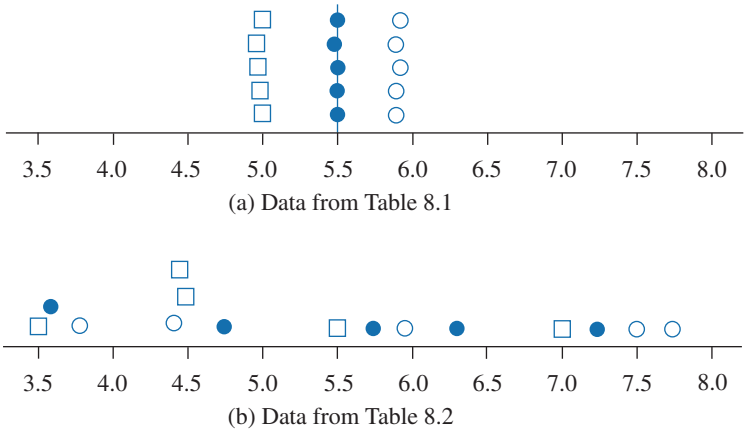


FIGURE 8.1
Dot diagrams for the data of Table 8.1 and Table 8.2: ○, measurement from sample 1; ●, measurement from sample 2; □, measurement from sample 3

means for the data of Table 8.1 is apparent in Figure 8.1(a). The lack of evidence to indicate a difference in population means for the data of Table 8.2 is indicated by the overlapping of data points for the samples in Figure 8.1(b).

analysis of variance

The preceding discussion, with the aid of Figure 8.1, should indicate what we mean by an **analysis of variance.** All differences in sample means are judged statistically significant (or not) by comparing them to the variation within samples. The details of the testing procedure will be presented after we discuss a research study that requires an analysis of variance to evaluate its research hypothesis.

Abstract of Research Study: Effect of Timing of the Treatment of Port-Wine Stains with Lasers

Port-wine stains are congenital vascular malformations that occur in an estimated 3 children per 1,000 births. The stigma of a disfiguring birthmark may have a substantial effect on a child's social and psychosocial adjustment. In 1985, the flash-pumped pulsed-dye laser was advocated for the treatment of port-wine stains in children. Treatment with this type of laser was hypothesized to be more effective in children than adults because the skin in children is thinner and the size of the port-wine stain is smaller; fewer treatments would therefore be necessary to achieve optimal clearance. These are all arguments to initiate treatment at an early age.

In a prospective study described in the paper ***"Effect of the Timing of Treatment of Port-Wine Stains with the Flash-Lamp-Pumped Pulsed-Dye Laser" (vander Horst et al., 1998),*** the researchers investigated whether treatment at a young age would yield better results than treatment at an older age.

One hundred patients, 31 years of age or younger, with a previously untreated port-wine stain were selected for inclusion in the study. During the first consultation, the extent and location of the port-wine stain were recorded. Four age groups of 25 patients each were determined for evaluating whether the laser treatment was more effective for younger patients.

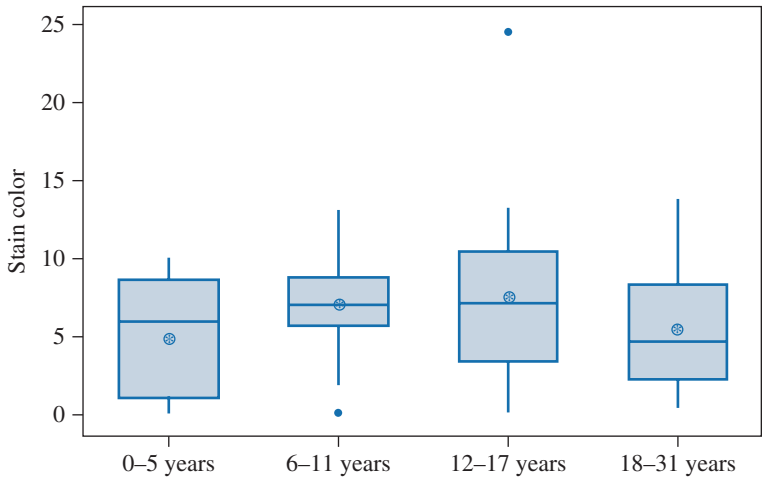
The summary statistics (Table 8.3) and boxplots (Figure 8.2) are provided for the four age groups. The 12–17 years group showed the greatest improvement, but the 6–11 years group had only a slightly smaller improvement. The other two groups had values at least two units less than the 12–17 years group. However, from the boxplots, we can observe that the four groups do not appear to have that

TABLE 8.3
Descriptive statistics for port-wine stain research study

Descriptive Statistics: 0-5 Years, 6-11 Years, 12-17 Years, 18-31 Years

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
0-5 Years	21	4.999	3.916	.144	1.143	6.110	8.852	10.325
6-11 Years	24	7.224	3.564	.188	5.804	7.182	8.933	13.408
12-17 Years	21	7.76	5.46	.11	3.53	7.32	10.64	24.72
18-31 Years	23	5.682	4.147	.504	2.320	4.865	8.429	14.036

FIGURE 8.2
Boxplot of stain color by age group (means are indicated by circles)



great a difference in their improvements. In the next section, we will develop the analysis of variance procedure to confirm whether or not a statistically significant difference exists among the four age groups.

8.2 A Statistical Test About More Than Two Population Means: An Analysis of Variance

In Chapter 6, we presented a method for testing the equality of two population means. We hypothesized two normal populations (1 and 2) with means denoted by μ_1 and μ_2, respectively, and a common variance σ^2. To test the null hypothesis that $\mu_1 = \mu_2$, independent random samples of sizes n_1 and n_2 were drawn from the two populations. The sample data were then used to compute the value of the test statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{(1/n_1) + (1/n_2)}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

pooled estimate of σ^2

is a **pooled estimate of** the common population variance σ^2**.** The rejection region for a specified value of α, the probability of a Type I error, was then found using Table 2 in the Appendix.

Now suppose that we wish to extend this method to test the equality of more than two population means. The test procedure described here applies to only

two means and therefore is inappropriate. Hence, we will employ a more general method of data analysis, the analysis of variance. We illustrate its use with the following example.

College students from five regions of the United States—northeast, southeast, midwest, southwest, and west—were interviewed to determine their attitudes toward industrial pollution. Each student selected was asked a set of questions related to the impact on economic development of proposed federal restrictions on air and water pollution. A total score reflecting each student's responses was then produced. Suppose that 250 students are randomly selected in each of the five regions. We wish to examine the average student score for each of the five regions.

We label the set of all test scores that could have been obtained from region I as population I, and we assume that this population possesses a mean μ_1. A random sample of $n_1 = 250$ measurements (scores) is obtained from this population to monitor student attitudes toward pollution. The set of all scores that could have been obtained from students from region II is labeled population II (which has a mean μ_2). A random sample of $n_2 = 250$ scores is obtained from this population. Similarly μ_3, μ_4, and μ_5 represent the means of the populations for scores from regions III, IV, and V, respectively. We also obtain random samples of 250 student scores from each of these populations.

From each of these five samples, we calculate a sample mean and variance. The sample results can then be summarized as shown in Table 8.4.

If we are interested in testing the equality of the population means (i.e., $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$), we might be tempted to run all possible pairwise comparisons of two population means. Hence, if we confirm that the five distributions are approximately normal with the same variance, σ^2, we could run 10 t tests comparing all pairs of means, as listed here (see Section 6.2).

Null Hypotheses

$\mu_1 = \mu_2$	$\mu_1 = \mu_4$	$\mu_2 = \mu_3$	$\mu_2 = \mu_5$	$\mu_3 = \mu_5$
$\mu_1 = \mu_3$	$\mu_1 = \mu_5$	$\mu_2 = \mu_4$	$\mu_3 = \mu_4$	$\mu_4 = \mu_5$

multiple *t* tests

One obvious disadvantage to this test procedure is that it is tedious and time consuming. However, a more important and less apparent disadvantage of running **multiple *t* tests** to compare means is that the probability of falsely rejecting at least one of the hypotheses increases as the number of t tests increases. Thus, although we may have the probability of a Type I error fixed at $\alpha = .05$ for each individual test, the probability of falsely rejecting *at least one* of those tests is larger than .05. In other words, the combined probability of a Type I error for the set of 10 hypotheses would be much larger than the value .05 set for each individual test. Indeed, it can be proved that the combined probability could be as large as .40.

TABLE 8.4
Summary of the sample results for five populations

	Population				
	I	II	III	IV	V
Sample mean	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$
Sample variance	s_1^2	s_2^2	s_3^2	s_4^2	s_5^2
Sample size	250	250	250	250	250

What we need is a single test of the hypothesis "all five population means are equal" that will be less tedious than the individual t tests and can be performed with a specified probability of a Type I error (say, .05). This test is the analysis of variance.

The analysis of variance procedures are developed under the following conditions:

1. Each of the five populations has a normal distribution.
2. The variances of the five populations are equal; that is, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma^2$.
3. The five sets of measurements are independent random samples from their respective populations.

From condition 2, we now consider the quantity

s_W^2

$$s_W^2 = \frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2 + (n_5 - 1)s_5^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + (n_4 - 1) + (n_5 - 1)}$$

$$= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2 + (n_5 - 1)s_5^2}{n_1 + n_2 + n_3 + n_4 + n_5 - 5}$$

Note that this quantity is merely an extension of

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

which is used as an estimate of the common variance for two populations for a test of the hypothesis $\mu_1 = \mu_2$ (Section 6.2). Thus, s_W^2 represents a combined estimate of the common variance σ^2, and it measures the variability of the observations within the five populations. (The subscript W refers to the within-sample variability.)

Next, we consider a quantity that measures the variability among the population means. If the null hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ is true, then the populations are identical, with mean μ and variance σ^2. Drawing single samples from the five populations is then equivalent to drawing five different samples from the same population. What kind of variation might we expect for these sample means? If the variation is too great, we would reject the hypothesis that $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$.

To evaluate the variation in the five sample means, we need to know the sampling distribution of the sample mean computed from a random sample of 250 observations from a normal population. From our discussion in Chapter 4, we recall that the sampling distribution for $\bar{y}$ based on $n = 250$ measurements will have the same mean as the population, μ, but the variance of $\bar{y}$ will be $\sigma^2/250$. We have five random samples of 250 observations each, so we can estimate the variance of the distribution of sample means, $\sigma^2/250$, using the formula

$$\text{Sample variance of five sample means} = \frac{\sum_{i=1}^{5}(\bar{y}_i - \bar{y}_\cdot)^2}{5 - 1}$$

where $\bar{y}_\cdot = \sum_{i=1}^{5} \bar{y}_i/5$ is the average of the five $\bar{y}_i$s.

Note that we merely consider the $\bar{y}$s to be a sample of five observations and calculate the "sample variance." This quantity estimates $\sigma^2/250$, and, hence, $250 \times$ (sample variance of the means) estimates σ^2. We designate this quantity as s_B^2; the subscript B denotes a measure of the variability among the sample means for the five populations. For this problem, $s_B^2 =$ (250 times the sample variance of the means).

s_B^2

Under the null hypothesis that all five population means are identical, we have two estimates of σ^2—namely, s_W^2 and s_B^2. Suppose the ratio

$$\frac{s_B^2}{s_W^2}$$

is used as the test statistic to test the hypothesis that $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. What is the distribution of this quantity if we repeat the experiment over and over again, each time calculating s_B^2 and s_W^2?

For our example, s_B^2/s_W^2 follows an F distribution with degrees of freedom that can be shown to be $df_1 = 4$ for s_B^2 and $df_2 = 1{,}245$ for s_W^2. The proof of these remarks is beyond the scope of this text. However, we will make use of this result for testing the null hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$.

test statistic

The **test statistic** used to test equality of the population means is

$$F = \frac{s_B^2}{s_W^2}$$

When the null hypothesis is true, both s_B^2 and s_W^2 estimate σ^2, and we expect F to assume a value near $F = 1$. When the hypothesis of equality is false, s_B^2 will tend to be larger than s_W^2 due to the differences among the population means. Hence, we will reject the null hypothesis in the upper tail of the distribution of $F = s_B^2/s_W^2$; for $\alpha = .05$, the critical value of $F = s_B^2/s_W^2$ is 2.37. (See Figure 8.3.) If the calculated value of F falls in the rejection region, we conclude that not all five population means are identical.

This procedure can be generalized (and simplified) with only slight modifications in the formulas to test the equality of t (where t is an integer equal to or greater than 2) population means from normal populations with a common variance σ^2. Random samples of sizes $n_1, n_2, \ldots, n_t$ are drawn from the respective populations. We then compute the sample means and variances. The null hypothesis $\mu_1 = \mu_2 = \cdots = \mu_t$ is tested against the alternative that at least one of the population means is different from the others.

Before presenting the generalized test procedure, we introduce the notation to be used in the formulas for s_B^2 and s_W^2.

completely randomized design

The experimental setting in which a random sample of observations is taken from each of t different populations is called a **completely randomized design.** Consider a completely randomized design in which four observations are obtained from each of the five populations. If we let y_{ij} denote the jth observation from population i, we could display the sample data for this completely randomized design as shown in Table 8.5. Using Table 8.5, we can introduce notation that is helpful when performing an **analysis of variance (AOV)** for a completely randomized design.

analysis of variance

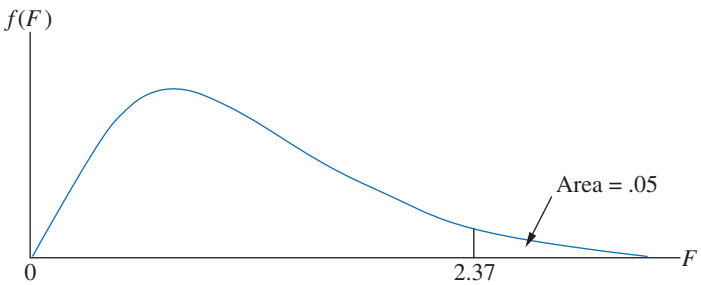


FIGURE 8.3
Critical value of F for $\alpha = .05$, $df_1 = 4$, and $df_2 = 1{,}245$

TABLE 8.5
Summary of sample data for a completely randomized design

Population	Data				Mean
1	y_{11}	y_{12}	y_{13}	y_{14}	$\bar{y}_{1.}$
2	y_{21}	y_{22}	y_{23}	y_{24}	$\bar{y}_{2.}$
3	y_{31}	y_{32}	y_{33}	y_{34}	$\bar{y}_{3.}$
4	y_{41}	y_{42}	y_{43}	y_{44}	$\bar{y}_{4.}$
5	y_{51}	y_{52}	y_{53}	y_{54}	$\bar{y}_{5.}$

Notation Needed for the AOV of a Completely Randomized Design

y_{ij}: The jth sample observation selected from population i. For example, y_{23} denotes the third sample observation drawn from population 2.

n_i: The number of sample observations selected from population i. In our data set, n_1, the number of observations obtained from population 1, is 4. Similarly, $n_2 = n_3 = n_4 = n_5 = 4$. However, it should be noted that the sample sizes need not be the same. Thus, we might have $n_1 = 12$, $n_2 = 3$, $n_3 = 6$, $n_4 = 10$, and so forth.

n_T: The total sample size; $n_T = \sum n_i$. For the data given in Table 8.5, $n_T = n_1 + n_2 + n_3 + n_4 + n_5 = 20$.

$\bar{y}_{i.}$: The average of the n_i sample observations drawn from population i; $\bar{y}_{i.} = \sum_j y_{ij}/n_i$.

$\bar{y}_{..}$: The average of all sample observations; $\bar{y}_{..} = \sum_i \sum_j y_{ij}/n_T$.

With this notation, it is possible to establish the following algebraic identities. (Although we will use these results in later calculations for s_W^2 and s_B^2, the proofs of these identities are beyond the scope of this text.) We can measure the variability of the n_T sample measurements y_{ij} about the overall mean $\bar{y}_{..}$ using the quantity

$$\text{TSS} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

total sum of squares

This quantity is called the **total sum of squares** (TSS) of the measurements about the overall mean. The double summation in TSS means that we must sum the squared deviations for all rows (i) and columns (j) of the one-way classification.

It is possible to partition the total sum of squares as follows:

$$\sum_{i,j} (y_{ij} - \bar{y}_{..})^2 = \sum_{i,j} (y_{ij} - \bar{y}_{i.})^2 + \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

The first quantity on the right side of the equation measures the variability of an observation y_{ij} about its sample mean $\bar{y}_{i.}$. Thus,

$$\text{SSW} = \sum_{i,j} (y_{ij} - \bar{y}_{i.})^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_t - 1)s_t^2$$

within-sample sum of squares

is a measure of the *within-sample* variability. SSW is referred to as the **within-sample sum of squares** and is used to compute s_W^2.

The second expression in the total sum of squares equation measures the variability of the sample means $\bar{y}_{i.}$ about the overall mean $\bar{y}_{..}$. This quantity, which

sum of squares between samples

measures the variability *between* (or among) the sample means, is referred to as the **sum of squares between samples** (SSB) and is used to compute s_B^2.

$$\text{SSB} = \sum_i n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

Although the formulas for TSS, SSW, and SSB are easily interpreted, they are not easy to use for calculations. Instead, we recommend using a computer software program.

An analysis of variance for a completely randomized design with t populations has the following null and alternative hypotheses:

H_0: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_t$ (i.e., the t population means are equal)

H_a: At least one of the t population means differs from the rest.

The quantities s_B^2 and s_W^2 can be computed using the shortcut formulas

$$s_B^2 = \frac{\text{SSB}}{t-1} \quad s_W^2 = \frac{\text{SSW}}{n_T - t}$$

where $t - 1$ and $n_T - t$ are the degrees of freedom for s_B^2 and s_W^2, respectively.

mean square

Historically, people have referred to a sum of squares divided by its degrees of freedom as a **mean square.** Hence, s_B^2 is often called the *mean square between samples* and s_W^2 the *mean square within samples.* The quantities are the mean squares because they both are averages of squared deviations. There are only $n_T - t$ linearly independent deviations $(y_{ij} - \bar{y}_{i.})$ in SSW because $\sum_j(y_{ij} - \bar{y}_{i.}) = 0$ for each of the t samples. Hence, we divide SSW by $n_T - t$ and not n_T. Similarly, there are only $t - 1$ linearly independent deviations $(\bar{y}_{i.} - \bar{y}_{..})$ in SSB because $\sum_i n_i(\bar{y}_{i.} - \bar{y}_{..}) = 0$. Hence, we divide SSB by $t - 1$.

The null hypothesis of equality of the t population means is rejected if

$$F = \frac{s_B^2}{s_W^2}$$

exceeds the tabulated value of F for the specified value of α, $\text{df}_1 = t - 1$, and $\text{df}_2 = n_T - t$.

AOV table

After we complete the F test, we then summarize the results of the study in an *analysis of variance table.* The format of an **AOV table** is shown in Table 8.6. The AOV table lists the sources of variability in the first column. The second column lists the sums of squares associated with each source of variability. We showed that the total sum of squares (TSS) can be partitioned into two parts, so SSB and SSW must add up to TSS in the AOV table. The third column of the table gives the degrees of freedom associated with the sources of variability. Again, we have a check; $(t - 1) + (n_T - t)$ must add up to $n_T - 1$. The mean squares are found in the fourth column of Table 8.6, and the F test for the equality of the t population means is given in the fifth column.

TABLE 8.6
An example of an AOV table for a completely randomized design

Source	Sum of Squares	Degrees of Freedom	Mean Square	*F* Test
Between samples	SSB	$t - 1$	$s_B^2 = \text{SSB}/(t-1)$	s_B^2/s_W^2
Within samples	SSW	$n_T - t$	$s_W^2 = \text{SSW}/(n_T - t)$	
Totals	TSS	$n_T - 1$		

EXAMPLE 8.1

A large body of evidence shows that soy has health benefits for most people. Some of these benefits come largely from isoflavones, plant compounds that have estrogen-like properties. The amount of isoflavones varies widely depending on the type of food processing. A consumer group purchased various soy products and ran laboratory tests to determine the amount of isoflavones in each product. There were three major categories of soy products: cereals and snacks (1), energy bars (2), and veggie burgers (3). Five different products from each of the three categories were selected, and the amount of isoflavones (in mg) was determined for an adult serving of the product. The consumer group wanted to determine if the average amount of isoflavones was different for the three sources of soy products. The data are given in Table 8.7. Use these data to test the research hypothesis of a difference in the mean isoflavone levels for the three categories. Use $\alpha = .05$.

TABLE 8.7
Isoflavone content from three sources of soy

Source of Soy	Isoflavone Content (mg)					Sample Sizes	Sample Means	Sample Variances
1	3	17	12	10	4	5	9.20	33.7000
2	19	10	9	7	5	5	10.00	29.0000
3	25	15	12	9	8	5	13.80	46.7000
Overall						15	11.00	

Solution The null and alternative hypotheses for this example are

H_0: $\mu_1 = \mu_2 = \mu_3$
H_a: At least one of the three population means is different from the rest.

The sample sizes are $n_1 = n_2 = n_3 = 5$, which yields $n_T = 15$. Using the sample means and sample variances, the sums of squares within and between are given here with

$$\bar{y}_{..} = (5\bar{y}_{1.} + 5\bar{y}_{2.} + 5\bar{y}_{3.})/15 = (5(9.20) + 5(10.00) + 5(13.80))/15 = 11.00$$

$$\text{SSB} = \sum_{i=1}^{3} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= 5(9.20 - 11.00)^2 + 5(10.00 - 11.00)^2 + 5(13.80 - 11.00)^2 = 60.40$$

and

$$\text{SSW} = \sum_{i=1}^{3}(n_i - 1)s_i^2 = (5 - 1)(33.7) + (5 - 1)(29.0) + (5 - 1)(46.7)$$

$$= 437.60$$

Finally, TSS = SSB + SSW = 60.40 + 437.60 = 498.00.

The AOV table for these data is shown in Table 8.8. The critical value of $F = s_B^2/s_W^2$ is 3.89, which is obtained from Table 8 in the Appendix for $\alpha = .05$, $df_1 = 2$, and $df_2 = 12$. Because the computed value of F, 0.83, does not exceed 3.89, we fail to reject the null hypothesis of equality of the mean levels of isoflavones for the three categories of soy products. Thus, there is not significant evidence that the three categories of soy products provide on the average different levels of isoflavones.

The p-value is computed to be $p\text{-value} = 1 - pf(.83, 2, 12) = .4596$.

TABLE 8.8
AOV Table for Example 8.1

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Test
Between samples	60.40	2	60.40/2 = 30.20	30.20/36.47 = 0.83
Within samples	437.60	12	437.60/12 = 36.47	
Total	498.00	14		

■

EXAMPLE 8.2

A clinical psychologist wished to compare three methods for reducing hostility levels in university students and used a certain test (HLT) to measure the degree of hostility. A high score on the test indicated great hostility. The psychologist used 24 students who obtained high and nearly equal scores in the experiment. Eight were selected at random from among the 24 problem cases and were treated with method 1. Seven of the remaining 16 students were selected at random and treated with method 2. The remaining nine students were treated with method 3. All treatments were continued for a one-semester period. Each student was given the HLT test at the end of the semester, with the results shown in Table 8.9. Use these data to perform an analysis of variance to determine whether there are differences among mean scores for the three methods. Use $\alpha = .05$.

TABLE 8.9
HLT test scores

Method	Test Scores									Mean	Standard Deviation	Sample Size
1	96	79	91	85	83	91	82	87		86.750	5.625	8
2	77	76	74	73	78	71	80			75.571	3.101	7
3	66	73	69	66	77	73	71	70	74	71.000	3.674	9

Solution The null and alternative hypotheses are

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: At least one of the population means differs from the rest.

For $n_1 = 8, n_2 = 7$, and $n_3 = 9$, we have a total sample size of $n_T = 24$. Using the sample means given in the table, we compute the overall mean of the 24 data values:

$$\bar{y}_{..} = \sum_{i=1}^{3} n_i \bar{y}_{i.}/n_T = (8(86.750) + 7(75.571) + 9(71.000))/24 = 1{,}861.997/24 = 77.5832$$

Using this value along with the means and standard deviations in Table 8.9, we can compute the three sums of squares as follows:

$$\text{SSB} = \sum_{i=1}^{3} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = 8(86.750 - 77.5832)^2 + 7(75.571 - 77.5832)^2 + 9(71 - 77.5832)^2 = 1{,}090.6311$$

and

$$\text{SSW} = \sum_{i=1}^{3}(n_i - 1)s_i^2 = (8 - 1)(5.625)^2 + (7 - 1)(3.101)^2 + (9 - 1)(3.674)^2$$
$$= 387.1678$$

Finally, TSS = SSB + SSW = 1,090.6311 + 387.1678 = 1,477.7989. The AOV table for these data is given in Table 8.10.

TABLE 8.10
AOV table for data of Example 8.2

Source	SS	df	MS	*F*	*p*-value
Between samples	1,090.6311	2	545.316	545.316/18.4366 = 29.58	<.001
Within samples	387.1678	21	18.4366		
Total	1,477.7989	23			

The critical value of F is obtained from Table 8 in the Appendix for $\alpha = .05$, $df_1 = 2$, and $df_2 = 21$; this value is 3.47. Because the computed value of F is 29.58, which exceeds the critical value 3.47, we reject the null hypothesis of equality of the mean scores for the three methods of treatment. The p-value is computed to be $p\text{-value} = 1 - pf(29.58, 2, 21) = .00000078$. Thus, there is a very strong rejection of the null hypothesis. From the three sample means, we observe that the mean for method 1 is considerably larger than the means for methods 2 and 3. The researcher would need to determine whether all three population means differ or whether the means for methods 2 and 3 are equal. Also, we may want to place confidence intervals on the three method means and on their differences; this would provide the researcher with information concerning the degree of differences in the three methods. In the next chapter, we will develop techniques to construct these types of inferences. Computer output shown next has slightly different values due to rounding in our manual calculations. In the computer printout, note that the names for the sum of squares are not given as between and within. The between sum of squares is labeled by Model. The within sum of squares is labeled as Error.

```
General Linear Models Procedure

Class Level Information

Class    Levels  Values
METHOD        3  1 2 3

Number of observations in data set = 24

Dependent Variable: SCORE

Source            DF   Sum of Squares   F Value   Pr > F
Model              2    1090.61904762     29.57   0.0001
Error             21     387.21428571
Corrected Total   23   1477.833333333
```

8.3 The Model for Observations in a Completely Randomized Design

In this section, we will consider a model for the completely randomized design (sometimes referred to as a one-way classification). This model will demonstrate the types of settings for which AOV testing procedures are appropriate. We can think of a model as a mathematical description of a physical setting. A model also enables us to computer-simulate the data that the physical process generates.

We will impose the following conditions concerning the sample measurements and the populations from which they are drawn:

1. The samples are independent random samples. Results from one sample in no way affect the measurements observed in another sample.
2. Each sample is selected from a normal population.
3. The mean and variance for population i are, respectively, μ_i and $\sigma_i^2 (i = 1, 2, \ldots, t)$. The t variances are equal: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 = \sigma^2$.

Figure 8.4 depicts a setting in which these three conditions are satisfied. The population distributions are normal with the same standard deviation. Note that populations III and IV have the same mean, which differs from the means of populations I and II. To summarize, we assume that the t populations are independently normally distributed with different means but a common variance σ^2.

We can now formulate a model (equation) that encompasses these three assumptions. Recall that we previously let y_{ij} denote the jth sample observation from population i.

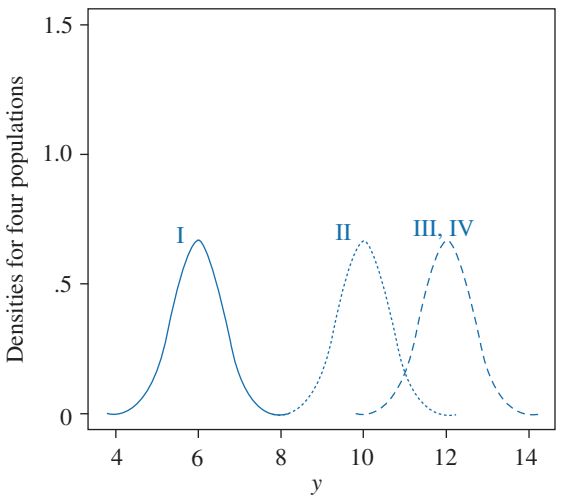
$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

An initial interpretation of the model will be given next. However, we will later explain why this interpretation needs to be modified in order to obtain appropriate estimators of the parameters using the observed data.

model terms

One interpretation of the **model** is that y_{ij}, the jth sample measurement selected from population i, is the sum of three **terms.** The term μ denotes the overall mean across all t populations—that is, the mean of the population consisting

FIGURE 8.4
Distributions of four populations that satisfy AOV assumptions



of the observations from all t populations. The term τ_i denotes the effect of population i on the differences in the t population means. The terms μ and τ_i are unknown constants, which will be estimated from the data obtained during the study or experiment. The term ε_{ij} represents the random deviation of y_{ij} about the ith population mean, μ_i. The ε_{ij}s are often referred to as *error terms.* The expression *error* is not to be interpreted as a mistake made in the experiment. Instead, the ε_{ij}s model the random variation of the y_{ij}s about their mean μ_i. The term *error* simply refers to the fact that the observations from the t populations differ by more than just their means. We assume the ε_{ij}s are independently normally distributed with a mean of 0 and a standard deviation of σ_ε. The independence condition can be interpreted as follows: The ε_{ij}s are independent if the size of the deviation of the y_{ij} observation from μ_i in no way affects the size of the deviation associated with any other observation.

Since y_{ij} is an observation from the ith population, it has mean μ_i. However, since the ε_{ij}s are distributed with mean 0, the mean or expected value of y_{ij}, denoted by $E(y_{ij})$, is

$$\mu_i = E(y_{ij}) = E(\mu + \tau_i + \varepsilon_{ij}) = \mu + \tau_i + E(\varepsilon_{ij}) = \mu + \tau_i$$

One problem with expressing the treatment means as $\mu_i = \mu + \tau_i$ is that we then have an overparameterized model. This occurs because there are only t treatment means, $\mu_1, \mu_2, \ldots, \mu_t$, but we have $t + 1$ parameters: μ and $\tau_1, \tau_2, \ldots, \tau_t$ in the model. In order to obtain the least squares estimates, it is necessary to put constraints on this set of parameters. A widely used constraint is to set $\tau_i = 0$. Then we have exactly t parameters in our description of the t treatment means. However, this results in the following interpretation of the parameters:

$$\mu = \mu_t \quad \tau_1 = \mu_1 - \mu_t, \quad \tau_2 = \mu_2 - \mu_t, \quad \cdots, \quad \tau_{t-1} = \mu_{t-1} - \mu_t, \quad \tau_t = 0$$

Thus, for $i = 1, 2, \cdots, t - 1$, τ_i is comparing μ_i to μ_t. This is the parametrization used by most software programs. The variance for each of the t populations are required to be σ_ε^2. Finally, because the εs are normally distributed, each of the t populations is normal. A summary of the assumptions for a one-way classification is shown in Table 8.11.

The null hypothesis for a one-way analysis of variance is that $\mu_1 = \mu_2 = \cdots = \mu_t$. Using our model, this would be equivalent to the null hypothesis

$$H_0\colon\ \tau_1 = \tau_2 = \cdots = \tau_t = 0$$

If H_0 is true, then all populations have the same unknown mean μ. Indeed, many textbooks use this latter null hypothesis for the analysis of variance in a completely randomized design. The corresponding alternative hypothesis is

$$H_a\colon\ \text{At least one of the } \tau_i\text{s differs from 0.}$$

TABLE 8.11
Summary of some of the assumptions for a completely randomized design

Population	Population Mean	Population Variance	Sample Measurements
1	$\mu + \tau_1$	σ_ε^2	$y_{11}, y_{12}, \ldots, y_{1n_1}$
2	$\mu + \tau_2$	σ_ε^2	$y_{21}, y_{22}, \ldots, y_{2n_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t	$\mu + \tau_t$	σ_ε^2	$y_{t1}, y_{t2}, \ldots, y_{tn_t}$

In this section, we have presented a brief description of the model associated with the analysis of variance for a completely randomized design. Although some authors bypass an examination of the model, we believe it is a necessary part of an analysis of variance discussion.

We have imposed several conditions on the populations from which the data are selected or, equivalently, on the experiments in which the data are generated, so we need to verify that these conditions are satisfied prior to making inferences from the AOV table. In Chapter 7, we discussed how to test the "equality of variances" condition using the BFL test. The normality condition is not as critical as the equal variance assumption when we have large sample sizes unless the populations are severely skewed or have very heavy tails. When we have small sample sizes, the normality condition and the equal variance condition become more critical. This situation presents a problem because there generally will not be enough observations from the individual population to test validly whether the normality or equal variance condition is satisfied. In the next section, we will discuss a technique that can at least partially overcome this problem. Also, some alternatives to the AOV will be presented in later sections of this chapter that can be used when the populations have unequal variances or have nonnormal distributions. As we discussed in Chapter 6, the most critical of the three conditions is that the data values are independent. This condition can be met by carefully conducting the studies or experiments so as to not obtain data values that are dependent. In studies involving randomly selecting data from the t populations, we need to take care that the samples are truly random and that the samples from one population are not dependent on the values obtained from another population. In experiments in which t treatments are randomly assigned to experimental units, we need to make sure that the treatments are truly **randomly assigned.** Also, the experiments must be conducted so the experimental units do not interact with each other in a manner that could affect their responses.

randomly assigned

8.4 Checking on the AOV Conditions

The assumption of equal population variances and the assumption of normality of the populations have been made in several places in the text, such as for the t test when comparing two population means and now for the analysis of variance F test in a completely randomized design.

Let us consider first an experiment in which we wish to compare t population means based on independent random samples from each of the populations. Recall that we assume we are dealing with normal populations with a common variance σ_ε^2 and possibly different means. We could verify the assumption of equality of the population variances using the BFL test of Chapter 7.

Several comments should be made here. Most practitioners do not routinely run a test of equality of variances. Fortunately, as we mentioned in Chapter 6, the assumption of homogeneity (equality) of population variances is less critical when the sample sizes are nearly equal; then the variances can be markedly different, and the p-values for an analysis of variance will still be only mildly distorted. In extreme situations, where homogeneity of the population variances is a problem, a transformation of the data may help to stabilize the variances.Then inferences can be made from an analysis of variance.

The normality of the population distributions can be checked using normal probability plots or boxplots, as we discussed in Chapters 5 and 6, when the

sample sizes are relatively large. However, in many experiments, the sample sizes may be as small as three to five observations from each population. In this case, the plots will not be a very reliable indication of whether the population distributions are normal. By taking into consideration the model we introduced in the previous section, the evaluation of the normal condition will be evaluated using a **residuals analysis.**

residuals analysis

From the model, we have $y_{ij} = \mu + \tau_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}$. Thus, we can write $\varepsilon_{ij} = y_{ij} - \mu_i$. Then, if the condition of equal variances is valid, the ε_{ij}s are a random sample from a normal population. However, μ_i is an unknown constant, but if we estimate μ_i with $\bar{y}_{i.}$ and let

$$e_{ij} = y_{ij} - \bar{y}_{i.}$$

then we can use the e_{ij}s to evaluate the normality assumption. Even when the individual n_is are small, we would have n_T residuals, which would provide a sufficient number of values to evaluate the normality condition. We can plot the e_{ij}s in a boxplot or a normal probability plot to evaluate whether the data appear to have been generated from normal populations.

EXAMPLE 8.3

Because many HMOs either do not cover mental health costs or provide only minimal coverage, ministers and priests often need to provide counseling to persons suffering from mental illness. An interdenominational organization wanted to determine whether the clerics from different religions have different levels of awareness with respect to the causes of mental illness. Three random samples were drawn, one containing 10 Methodist ministers, a second containing 10 Catholic priests, and a third containing 10 Pentecostal ministers. Each of the 30 clerics was then examined, using a standard written test, to measure his or her knowledge about causes of mental illness. The test scores are listed in Table 8.12. Does there appear to be a significant difference in the mean test scores for the three religions?

TABLE 8.12
Scores for clerics' knowledge of mental illness

Cleric	Methodist	Catholic	Pentecostal
1	62	62	37
2	60	62	31
3	60	24	15
4	25	24	15
5	24	22	14
6	23	20	14
7	20	19	14
8	13	10	5
9	12	8	3
10	6	8	2
$\bar{y}_{i.}$	30.50	25.90	15.00
s_i	21.66	20.01	11.33
n_i	10	10	10
Median($\tilde{y}_i$)	23.5	21	14

Solution Prior to conducting an AOV test of the three means, we need to evaluate whether the conditions required for AOV are satisfied. Figure 8.5 is a boxplot of the mental illness scores by religion. There is an indication that the data may be somewhat skewed to the right. Thus, we will evaluate the normality condition. We need to obtain the residuals $e_{ij} = y_{ij} - \bar{y}_{i.}$. For example, $e_{11} = y_{11} - \bar{y}_{1.} = 62 - 30.50 = 31.50$. The remaining e_{ij}s are given in Table 8.13.

FIGURE 8.5
Boxplots of awareness score (means are indicated by circles)

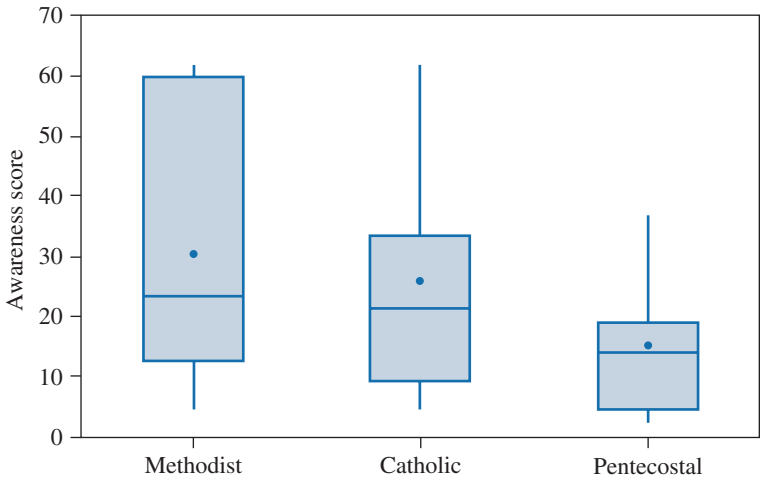


TABLE 8.13
Residuals e_{ij} for clerics' knowledge of mental illness

Cleric	Methodist	Catholic	Pentecostal
1	31.5	36.1	22.0
2	29.5	36.1	16.0
3	29.5	−1.9	0.0
4	−5.5	−1.9	0.0
5	−6.5	−3.9	−1.0
6	−7.5	−5.9	−1.0
7	−10.5	−6.9	−1.0
8	−17.5	−15.9	−10.0
9	−18.5	−17.9	−12.0
10	−24.5	−17.9	−13.0

The residuals are then plotted in Figures 8.6 and 8.7. The boxplot in Figure 8.7 displays three outliers out of 30 residuals. It is very unlikely that 10% of the data values are outliers if the residuals are in fact a random sample from a normal distribution. This is confirmed in the normal probability plot displayed in Figure 8.6, which shows a lack of concentration of the residuals about the straight line. Furthermore, the test of normality has a p-value less than .001, which indicates a strong departure from normality. Thus, we conclude that the data have nonnormal characteristics. In Section 8.6, we will provide an alternative to the F test from the AOV table, the Kruskal–Wallis test, which would be appropriate for this situation.

The BFL test is conducted to check the condition of equality of the variances in the three populations. An examination of the formula for the BFL test reveals that once we make the conversion of the data from y_{ij} to $z_{ij} = |y_{ij} - \tilde{y}_i|$, where $\tilde{y}_i$ is the sample median of the ith data set, the BFL test is equivalent to the F test

FIGURE 8.6
Normal probability plot for residuals

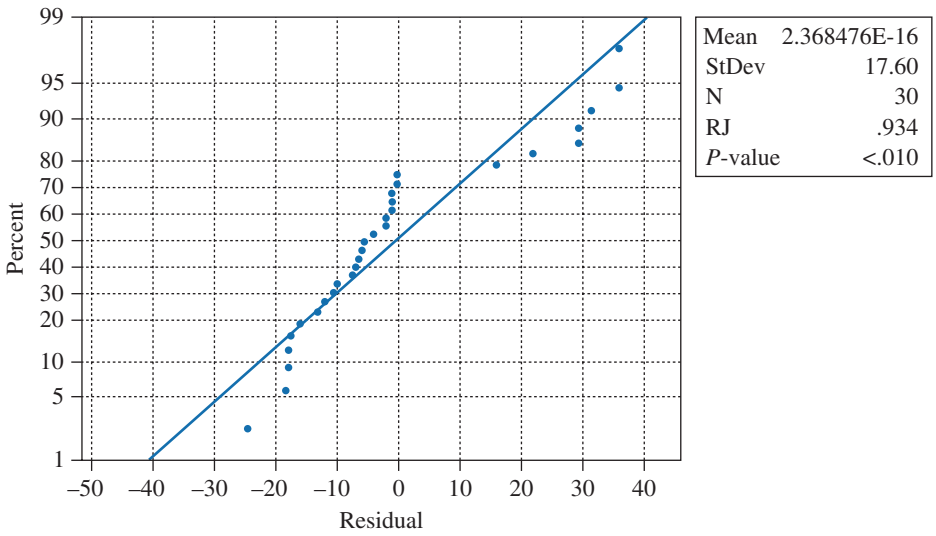
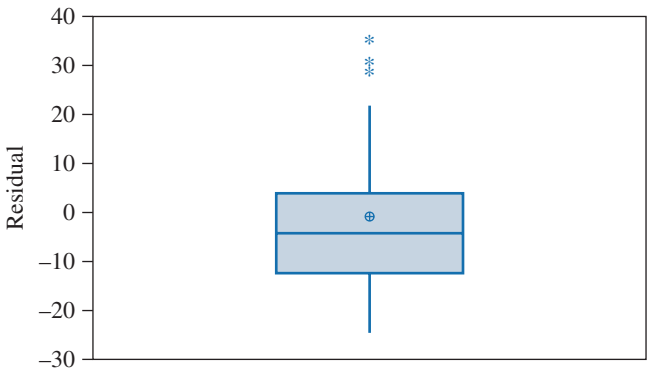


FIGURE 8.7
Boxplot of residuals



from AOV applied to the z_{ij}s. Thus, we can simply use the formulas from AOV to compute the BFL test. The z_{ij}s are given in Table 8.14 using the medians from Table 8.12.

TABLE 8.14
Transformed data set, $z_{ij} = |y_{ij} - \tilde{y}_i|$

Cleric	Methodist	Catholic	Pentecostal
1	38.5	41	23
2	36.5	41	17
3	36.5	3	1
4	1.5	3	1
5	0.5	1	0
6	0.5	1	0
7	3.5	2	0
8	10.5	11	9
9	11.5	13	11
10	17.5	13	12
$\bar{z}_{i.}$	15.70	12.90	7.40
s_i	15.80	15.57	8.29

Using the sample means given in the table, we compute the overall mean of the 30 data values:

$$\bar{z}_{..} = \sum_{i=1}^{3} n_i \bar{z}_{i.}/n_T = [10(15.70) + 10(12.90) + 10(7.40)]/30 = 360/30 = 12$$

Using this value along with the means and standard deviations in Table 8.14, we can compute the sum of squares as follows:

$$\text{SSB} = \sum_{i=1}^{3} n_i(\bar{z}_{i.} - \bar{z}_{..})^2 = 10(15.70 - 12)^2 + 10(12.90 - 12)^2 + 10(7.40 - 12)^2$$
$$= 356.6$$

and

$$\text{SSW} = \sum_{i=1}^{3} (n_i - 1)s_i^2 = (10 - 1)(15.80)^2 + (10 - 1)(15.57)^2$$
$$+ (10 - 1)(8.29)^2 = 5{,}047.10$$

The mean squares are MSB = SSB/$(t - 1)$ = 356.6/$(3 - 1)$ = 178.3 and MSW = SSW/$(n_T - t)$ = 5,047.10/$(30 - 3)$ = 186.9. Finally, we can next obtain the value of the BFL test statistic from L = MSB/MSW = 178.3/186.9 = .95. The critical value of L, using $\alpha = .05$, is obtained from the F tables with $df_1 = 2$ and $df_2 = 27$. This value is 3.35, and, thus, we fail to reject the null hypothesis that the standard deviations are equal. The p-value is greater than .25 because the smallest value in the F table with $df_1 = 2$ and $df_2 = 27$ is 1.46, which corresponds to a probability of 0.25. In fact, p-value = $1 - pf(.95, 2, 27) = .399$. Thus, we have a high degree of confidence that the three populations have the same variance. ■

In Section 8.6, we will present the Kruskal–Wallis test, which can be used when the populations are nonnormal but have identical distributions under the null hypothesis. This test requires, as a minimum, that the populations have the same variance. Thus, the Kruskal–Wallis test would not be appropriate for the situation in which the populations have very different variances. The next section will provide procedures for testing for differences in population means when the population variances are unequal.

8.5 An Alternative Analysis: Transformations of the Data

transformation of data

A **transformation of the sample data** is defined to be a process in which the measurements on the original scale are systematically converted to a new scale of measurement. For example, if the original variable is y and the variances associated with the variable across the treatments are not equal (heterogeneous), it may be necessary to work with a new variable such as $\sqrt{y}$, $\log y$, or some other transformed variable.

How can we select the appropriate transformation? This is no easy task and often takes a great deal of experience in the experimenter's area of application.

TABLE 8.15
Transformation to achieve uniform variance

Relationship Between μ and σ^2	y_T	Variance of y_T (for a given k)
$\sigma^2 = k\mu$ (when $k = 1$, y may be Poisson variable)	$y_T = \sqrt{y}$ or $\sqrt{y + .375}$	$1/4; (k = 1)$
$\sigma^2 = k\mu^2$	$y_T = \log y$ or $\log(y + 1)$	$1; (k = 1)$
$\sigma^2 = k\pi(1 - \pi)$ (when $k = 1/n$, y may be binomial variable)	$y_T = \sin^{-1}(\sqrt{y})$	$1/4n; (k = 1/n)$

guidelines for selecting y_T

In spite of these difficulties, we can consider several **guidelines for choosing an appropriate transformation.**

Many times the variances across the populations of interest are heterogeneous and seem to vary with the magnitude of the population mean. For example, it may be that the larger the population mean, the larger the population variance. When we are able to identify how the variance varies with the population mean, we can define a suitable transformation from the variable y to a new variable y_T. Three specific situations are presented in Table 8.15.

The first row of Table 8.15 suggests that if y is a Poisson* random variable, the variance of y is equal to the mean of y. Thus, if the different populations correspond to different Poisson populations, the variances will be heterogeneous provided the means are different. The transformation that will stabilize the variances is $y_T = \sqrt{y}$. However, if the Poisson means are small (under 5), the transformation $y_T = \sqrt{y + .375}$ is better.

EXAMPLE 8.4

Marine biologists are studying a major reduction in the number of shrimp and commercial fish in the Gulf of Mexico. The area in which the Mississippi River enters the gulf is one of the areas of greatest concern. The biologists hypothesize that nutrient-rich water, including mainly nitrogens from the farmlands of the Midwest, flows into the gulf, which results in rapid growth in algae that feeds zooplankton. Bacteria then feed on the zooplankton pellets and dead algae, resulting in a depletion of the oxygen in the water. The more mobile marine life flees these regions, while the less mobile marine life dies from hypoxia. To monitor this condition, the mean dissolved oxygen contents (in ppm) of four areas at increasing distance from the mouth of the Mississippi were determined. A random sample of 10 water samples was taken at a depth of 12 meters in each of the four areas. The sample data are given in Table 8.16. The biologists want to test whether the mean oxygen content is lower in those areas closer to the mouth of the Mississippi.

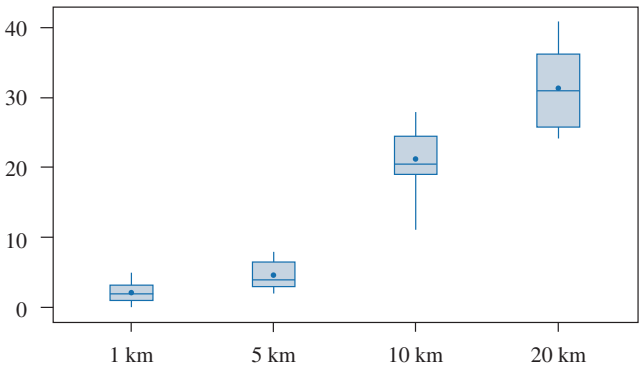
a. Run a test of the equality of the population variances with $\alpha = .05$.
b. Transform the data if necessary to obtain a new data set in which the observations have equal variances.

* The Poisson random variable was introduced in Chapter 4.

TABLE 8.16
Mean dissolved oxygen contents (in ppm) at four distances from mouth

Sample	Distance to Mouth			
	1 km	5 km	10 km	20 km
1	1	4	20	37
2	5	8	26	30
3	2	2	24	26
4	1	3	11	24
5	2	8	28	41
6	2	5	20	25
7	4	6	19	36
8	3	4	19	31
9	0	3	21	31
10	2	3	24	33
Mean	$\bar{y}_{1.} = 2.2$	$\bar{y}_{2.} = 4.6$	$\bar{y}_{3.} = 21.2$	$\bar{y}_{4.} = 31.4$
Standard Deviation	$s_1 = 1.476$	$s_2 = 2.119$	$s_3 = 4.733$	$s_4 = 5.5220$

FIGURE 8.8
Boxplots of 1–20 km (means are indicated by solid circles)



Solution

a. Figure 8.8 depicts the data in a set of boxplots. The data do not appear noticeably skewed or heavy-tailed. The BFL test is applied to the data and yields $L = 3.70$ with a p-value of .0203. This implies strong evidence of a difference in the four population variances.

b. We next examine the relationship between the sample means $\bar{y}_{i.}$ and sample variances s_i^2.

$$\frac{s_1^2}{\bar{y}_{1.}} = .99 \quad \frac{s_2^2}{\bar{y}_{2.}} = .97 \quad \frac{s_3^2}{\bar{y}_{3.}} = 1.06 \quad \frac{s_4^2}{\bar{y}_{4.}} = .97$$

Thus, it would appear that $\sigma_i^2 = k\mu_i$ with $k \approx 1$. From Table 8.15, the suggested transformation is $y_T = \sqrt{y + .375}$. The values of y_T appear in Table 8.17 along with their means and standard deviations. Although the original data had heterogeneous variances, the sample variances are all approximately .25, as indicated in Table 8.17.

TABLE 8.17
Transformation of data in Table 8.16: $y_T = \sqrt{y + .375}$

	Distance to Mouth			
Sample	**1 km**	**5 km**	**10 km**	**20 km**
1	1.173	2.092	4.514	6.114
2	2.318	2.894	5.136	5.511
3	1.541	1.541	4.937	5.136
4	1.173	1.837	3.373	4.937
5	1.541	2.894	5.327	6.432
6	1.541	2.318	4.514	5.037
7	2.092	2.525	4.402	6.031
8	1.837	2.092	4.402	5.601
9	0.612	1.837	4.623	5.601
10	1.541	1.837	4.937	5.777
Mean	1.54	2.19	4.62	5.62
Variance	.24	.22	.29	.24

■

coefficient of variation

The second transformation indicated in Table 8.15 ($\sigma^2 = k\mu^2$) is for an experimental situation in which the population variance is proportional to the square of the population mean or, equivalently, where $\sigma = \mu$. That is, the logarithmic transformation is appropriate any time the **coefficient of variation** σ_i/μ_i is constant across the populations of interest.

EXAMPLE 8.5

Arthritis is a very commonly occurring affliction. It is a major cause of lost work time and often results in serious disability. Of the many types of arthritis, the most common type is osteoarthritis. This condition is frequently due to wear and tear in the joints and is more likely to be found in people over 50. It is very painful in the weight-bearing joints, such as the knees and hips. Cartilage wears away on the bone ends, causing pain and swelling. Osteoarthritis may develop after an injury such as a bone fracture or a joint dislocation. In order to reduce the amount of time osteoarthritis patients are absent from work, it is important for them to have effective pain relief. An experiment was conducted to compare the effectiveness of three new analgesics: A_1, A_2, and A_3. A clinic evaluated a large group of patients and identified 45 patients with a moderate level of pain. Fifteen of the 45 persons were then randomly assigned to one of the three analgesics. The patients were then placed on the therapies, and the percentage reduction in pain level was assessed for each patient. These values are recorded in Table 8.18.

a. Are there significant differences among the population variances for the three analgesics? Use $\alpha = .05$.
b. Does it appear that the coefficient of variation is constant across the three therapies? If yes, then apply the log transformation to the data to try to stabilize the variances.
c. Compute the sample means and sample standard deviations for the transformed data. Did the transformation yield a stabilization of the variances?

TABLE 8.18
Data for percent reduction in pain

Subject	A_1	A_2	A_3
1	3.0	1.8	1.3
2	1.2	6.3	12.6
3	1.0	5.2	10.0
4	0.7	3.7	10.5
5	1.1	5.4	10.8
6	0.6	2.9	5.9
7	1.2	6.0	12.1
8	0.1	0.3	0.6
9	0.7	3.6	18.6
10	1.9	9.3	18.7
11	0.6	2.8	5.5
12	0.0	0.0	0.0
13	1.6	8.1	18.2
14	4.0	19.9	22.3
15	0.1	0.3	0.6
Mean $(\bar{y}_i)$	1.19	5.04	9.85
St. Dev.(s_i)	1.097	4.97	7.41
CV$(s_i/\bar{y}_i)$	.93	.99	.75

Solution

a. The BFL test for the hypothesis H_0: $\sigma^2_{A_1} = \sigma^2_{A_2} = \sigma^2_{A_3}$ was computed using Minitab. The results are given here. $L = 9.17$ with p-value $= 1 - pf(9.17, 2, 42) = .000496$. Thus, we reject H_0 and conclude that the populations variances are different.

b. The coefficients of variation (CV) for the three analgesics are very nearly the same value; thus we will apply the log transformation to the data. The transformed data are shown in Table 8.19. *Note:* Because there are 0s in the data, the transformation $y_T = \log(y + 1)$ should be computed. These values are shown in Table 8.19.

TABLE 8.19
Natural logarithms of the data in Table 8.18

Subject	A_1	A_2	A_3
1	1.38629	1.02962	.83291
2	.78846	1.98787	2.61007
3	.69315	1.82455	2.39790
4	.53063	1.54756	2.44235
5	.74194	1.85630	2.46810
6	.47000	1.36098	1.93152
7	.78846	1.94591	2.57261
8	.09531	.26236	.47000
9	.53063	1.52606	2.97553
10	1.06471	2.33214	2.98062
11	.47000	1.33500	1.87180
12	.00000	.00000	.00000
13	.95551	2.20827	2.95491
14	1.60944	3.03975	3.14845
15	.09531	.26236	.47000
Mean $(\bar{y}_i)$	.681	1.501	2.008
St. Dev.(s_i)	.455	.837	1.052

c. The sample means and standard deviations for the transformed data are given in Table 8.19. The BFL test for the transformed data yields $L = 2.207$ with a p-value $= 1 - pf(2.207, 2, 42) = .122$. Thus, we fail to reject H_0 and conclude that the transformation has produced data in which the three populations variances are approximately equal.

In Exercise 8.22, you will be asked to run an AOV test for differences in the mean pain reduction for both the transformed and the untransformed data to determine if the transformation resulted in a different conclusion concerning the effectiveness of the three analgesics. ■

$y_T = \arcsin \sqrt{y}$

The third transformation listed in Table 8.15 ($y_T = \arcsin \sqrt{y}$) is particularly appropriate for data recorded as percentages or proportions. Recall that in Chapter 4 we introduced the binomial distribution, where y designates the number of successes in n identical trials and $\hat{\pi} = y/n$ provides an estimate of π, the proportion of experimental units in the population possessing the characteristic. In Chapter 4, the variance of $\hat{\pi}$ was given by $\pi(1 - \pi)/n$. Thus, if the response variable is $\hat{\pi}$, the proportion of successes in a random sample of n observations, then the variance of $\hat{\pi}$ will vary depending on the values of π for the populations from which the samples were drawn. See Table 8.20.

From Table 8.20, we observe that the variance of $\hat{\pi}$ is symmetrical about $\pi = .5$. That is, the variance of $\hat{\pi}$ for $\pi = .7$ and $n = 20$ is .0105, the same value as for $\pi = .3$. The important thing to note is that if the populations have values of π in the vicinity of approximately .3 to .7, there is very little difference in the variances for $\hat{\pi}$. However, the variance of $\hat{\pi}$ is quite variable for either large or small values of π, and for these situations, we should consider the possibility of transforming the sample proportions to stabilize the variances.

The transformation we recommend is arcsin $\sqrt{\hat{\pi}}$ sometimes written as $\sin^{-1}(\sqrt{\hat{\pi}})$; that is, we are transforming the sample proportion into the angle whose sine is $\sqrt{\hat{\pi}}$. Some experimenters express these angles in degrees, others in radians. For consistency, we will always express our angles in radians. Table 9 of the Appendix provides arcsin computations for various values of $\hat{\pi}$.

TABLE 8.20
Variance of $\hat{\pi}$, the sample proportion, for several values of π and $n = 20$

Values of π	$\pi(1 - \pi)/n$	Values of π	$\pi(1 - \pi)/n$
.01	.0005	.99	.0005
.05	.0024	.95	.0024
.1	.0045	.90	.0045
.2	.0080	.80	.0080
.3	.0105	.70	.0105
.4	.0120	.60	.0120
.5	.0125		

EXAMPLE 8.6

A political action group conducted a national opinion poll to evaluate the voting public's opinion concerning whether the new EPA regulations on air pollution were stringent enough to protect the public's health. The group was also interested in determining if there were regional differences in the public's opinion concerning air pollution. For this poll, the country was divided into

four geographical regions (NE, SE, MW, W). A random sample of 100 registered voters was obtained from each of six standard metropolitan statistical areas (SMSAs) located in each of the four regions. The data in Table 8.21 are the sample proportions, $\hat{\pi}$, of people who thought the EPA standards were not stringent enough for the 24 SMSAs.

a. Is there a significant difference in the variability of the four region's proportion? Use $\alpha = .05$.
b. Transform the data using $y_T = \arcsin\sqrt{\hat{\pi}}$.
c. Compute the sample means and sample standard deviations for the transformed data. Did the transformation yield a stabilization of the variances?

TABLE 8.21
Sample proportions for the four regions

	Region			
SMSA	NE	SE	MW	W
1	.84	.43	.57	.10
2	.81	.35	.59	.12
3	.78	.27	.63	.13
4	.85	.40	.60	.15
5	.85	.28	.56	.11
6	.83	.33	.56	.11
Mean	.827	.343	.585	.120
Standard Deviation	.0273	.0638	.0274	.0179

Solution

a. The BFL test for the hypothesis H_0: $\sigma^2_{NE} = \sigma^2_{SE} = \sigma^2_{MW} = \sigma^2_W$ was computed to be $L = 3.55$ with p-value $= .033$. Thus, we reject H_0 and conclude that at the $\alpha = .05$ level there is significant evidence of a difference in the population variances.
b. Using a calculator, computer spreadsheet, or Table 9 in the Appendix, the transformed data are shown in Table 8.22.

TABLE 8.22
Arcsin of the square root of the sample proportions

	Region			
SMSA	NE	SE	MW	W
1	1.1593	.71517	.85563	.32175
2	1.1198	.63305	.87589	.35374
3	1.0826	.54640	.91691	.36886
4	1.1731	.68472	.88608	.39770
5	1.1731	.55760	.84554	.33807
6	1.1458	.61194	.84554	.33807
Mean	1.142	.625	.871	.353
Standard Deviation	.0354	.0673	.0279	.0271

c. We can observe that the standard deviations are more nearly alike than the standard deviations for the untransformed data. The BFL test for the hypothesis H_0: $\sigma^2_{NE} = \sigma^2_{SE} = \sigma^2_{MW} = \sigma^2_W$ was computed for the transformed data. The results are given here. $L = 2.44$ with p-value $= .094$. Thus, we fail to reject H_0 and conclude that at the $\alpha = .05$ level there is not significant evidence of a difference in the variances of the transformed proportions. ■

when $\pi = 0, 1$

One comment should be made concerning the situation in which a **sample proportion of 0 or 1** is observed. For these cases, we recommend substituting $1/4n$ and $1 - (1/4n)$, respectively, as the corresponding sample proportions to be used in the calculations.

power transformation

A general procedure for determining the appropriate transformation to stabilize the variances for the t treatment groups is the **power transformation.** The power transformation is discussed in the article ***"An Analysis of Transformations" (Box and Cox, 1964).*** The transformation is given by

$$y_T = \begin{cases} y^\lambda & \text{for } \lambda \neq 0 \\ \log y & \text{for } \lambda = 0 \end{cases}$$

The transformation includes as special cases the square root transformation, $\lambda = \frac{1}{2}$, and the natural logarithm, $\lambda = 0$. The Box–Cox method describes how to use the data to select the value of λ such that the transformed data more nearly meet the requirements of constant variance and normality. The book *Applied Regression Analysis* by Draper and Smith (1998) discusses in detail the Box–Cox family of transformations. This topic is also discussed in Chapter 13 when dealing with multiple regression.

In this section, we have discussed how transformations of data can alleviate the problem of nonconstant variances prior to conducting an analysis of variance. As an added benefit, the transformations presented in this section also (sometimes) decrease the nonnormality of the data. Still, there will be times when the presence of severe skewness or outliers causes nonnormality that could not be eliminated by these transformations. Wilcoxon's rank sum test (Chapter 6) can be used for comparing two populations in the presence of nonnormality when working with two independent samples. For data based on more than two independent samples, we can address nonnormality using the Kruskal–Wallis test (Section 8.6). Note that these tests are also based on a transformation (the rank transformation) of the sample data.

8.6 A Nonparametric Alternative: The Kruskal–Wallis Test

In Chapter 6, we introduced the Wilcoxon rank sum test for comparing two nonnormal populations. In this section, the rank sum test will be extended to a comparison of more than two populations. In particular, suppose that n_1 observations are drawn at random from population 1, n_2 from population 2, . . . , and n_k from population k. We may wish to test the hypothesis that the k samples were drawn from identical distributions. The following test procedure, sometimes called the Kruskal–Wallis test, is then appropriate.

Extension of the Rank Sum Test for More Than Two Populations

H_0: The k distributions are identical.
H_a: Not all the distributions are the same.

T.S.: $$H = \frac{12}{n_T(n_T+1)}\sum_i \frac{T_i^2}{n_i} - 3(n_T+1)$$

where n_i is the number of observations from sample i ($i = 1, 2, \ldots, k$), n_T is the combined (total) sample size (i.e., $n_T = \sum_i n_i$), and T_i denotes the sum of the ranks for the measurements in sample i after the combined sample measurements have been ranked

R.R.: For a specified value of α, reject H_0 if H exceeds the critical value of χ^2 for α and df = $k - 1$.

Note: When there are a large number of ties in the ranks of the sample measurements, use

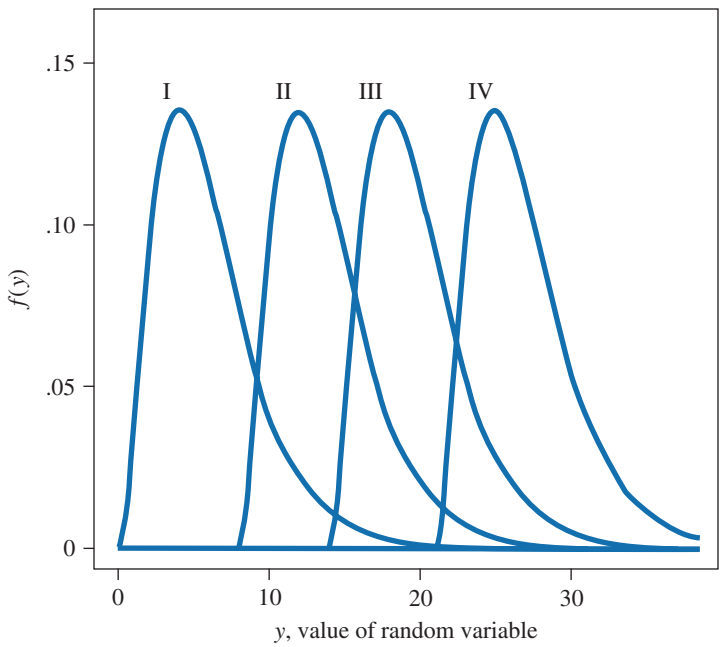
H′

$$H' = \frac{H}{1 - \left[\sum_j (t_j^3 - t_j)/(n_T^3 - n_T)\right]}$$

where t_j is the number of observations in the jth group of tied ranks.

Figure 8.9 displays population distributions under the alternative hypotheses of the Kruskal–Wallis test.

FIGURE 8.9
Four skewed population distributions identical in shape but shifted



EXAMPLE 8.7

Refer to Example 8.3, where we determined that the clerics' test scores were not normally distributed. Thus, we will apply the Kruskal–Wallis test to the data set displayed in Table 8.12.

Use the data to determine whether the three groups of clerics differ with respect to their knowledge about the causes of mental illness. Use $\alpha = .05$.

Solution The research and null hypotheses for this example can be stated as follows:

H_a: At least one of the three groups of clerics differs from the others with respect to knowledge about causes of mental illness.

H_0: There is no difference among the three groups with respect to knowledge about the causes of mental illness (i.e., the samples of scores were drawn from identical populations).

Before computing H, we must jointly rank the 30 test scores from lowest to highest. From Table 8.23, we see that 2 is the lowest test score, so we assign this cleric the rank of 1. Similarly, we give the scores 3, 5, and 6 the ranks 2, 3, and 4, respectively. Two clerics have a test score of 8, and because these two scores occupy the ranks 5 and 6, we assign each one a rank of 5.5—the average of the ranks 5 and 6. In a similar fashion, we can assign the remaining ranks to the test scores. Table 8.23 lists the 30 test scores and associated ranks (in parentheses).

TABLE 8.23
Scores for clerics' knowledge of mental illness, Example 8.3

Cleric	Methodist	Catholic	Pentecostal
1	62 (29)	62 (29)	37 (25)
2	60 (26.5)	62 (29)	31 (24)
3	60 (26.5)	24 (21)	15 (13.5)
4	25 (23)	24 (21)	15 (13.5)
5	24 (21)	22 (18)	14 (11)
6	23 (19)	20 (16.5)	14 (11)
7	20 (16.5)	19 (15)	14 (11)
8	13 (9)	10 (7)	5 (3)
9	12 (8)	8 (5.5)	3 (2)
10	6 (4)	8 (5.5)	2 (1)
Sum of ranks	182.5	167.5	115

Note from Table 8.23 that the sums of the ranks for the three groups of clerics are 182.5, 167.5, and 115. Hence, the computed value of H is

$$H = \frac{12}{30(30+1)}\left(\frac{(182.5)^2}{10} + \frac{(167.5)^2}{10} + \frac{(115)^2}{10}\right) - 3(30+1)$$
$$= \frac{12}{930}(3{,}330.625 + 2{,}805.625 + 1{,}322.5) - 93 = 3.24$$

Because there are groups of tied ranks, we will use H' and compare its value to H. To do this, we form the 20 groups composed of identical ranks, shown in Table 8.24.

From this information, we calculate the quantity

$$\sum_i \frac{(t_i^3 - t_i)}{n_T^3 - n_T}$$
$$= \frac{(2^3-2) + (3^3-3) + (2^3-2) + (2^3-2) + (3^3-3) + (2^3-2) + (3^3-3)}{30^3 - 30}$$
$$= .0036$$

TABLE 8.24
Ranked cleric data for Example 8.7

Rank	Group	t_i	Rank	Group	t_i
1	1	1	15	11	1
2	2	1	16.5, 16.5	12	2
3	3	1	18	13	1
4	4	1	19	14	1
5.5, 5.5	5	2	21, 21, 21	15	3
7	6	1	23	16	1
8	7	1	24	17	1
9	8	1	25	18	1
11, 11, 11	9	3	26.5, 26.5	19	2
13.5, 13.5	10	2	29, 29, 29	20	3

Substituting this value into the formula for H', we have

$$H' = \frac{H}{1-.0036} = \frac{3.24}{.9964} = 3.25$$

Thus, even with more than half of the measurements involved in ties, H' and H are nearly the same value. The critical value of the chi-square with $\alpha = .05$ and df $= k - 1 = 2$ can be found using Table 7 in the Appendix. This value is 5.991; we fail to reject the null hypothesis and conclude that there is no significant difference in the test scores of the three groups of clerics. It is interesting to note that the p-value for the Kruskal–Wallis test is $1 - pchisq(3.25, 2) = .197$, whereas the p-value from the AOV F test applied to the original test scores was .168. Thus, even though the data did not have a normal distribution, the F test from AOV is robust against departures from normality. Only when the data are extremely skewed or very heavily tailed do the Kruskal–Wallis test and the F test from AOV differ. ■

8.7 RESEARCH STUDY: Effect of Timing on the Treatment of Port-Wine Stains with Lasers

As was discussed at the beginning of this chapter, port-wine stains are disfiguring birthmarks that can be treated with a flash-pumped pulsed-dye laser. However, physicians wanted to investigate which age was the most effective time to administer the treatment. Younger patients tend to have thinner skin and smaller lesions, which may lead to a more effective treatment by the laser. A previous study found better results with early treatment, but the results were not unequivocally confirmed by a large number of similar studies. However, all of the studies were retrospective in nature, and in none of the studies were objective measurements used to assess the results.

Defining the Problem

Therefore, it was determined that a prospective study was needed to assess whether treatment of a port-wine stain at a young age would yield better results than treatment with older patients. Furthermore, an objective measurement of the reduction in the difference in color between skin with the stain and the contralateral healthy skin would need to be developed. In the paper ***"Effect of the Timing of Treatment of Port-Wine Stains with the Flash-Lamp-Pumped Pulsed-Dye Laser"***

(vander Horst et al., 1998), the researchers considered the following issues relative to the most effective treatment:

1. What objective measurements should be used to assess the effectiveness of the treatment in reducing the visibility of the port-wine stains?
2. How many different age groups should be considered for evaluating the treatment?
3. What type of experimental design would produce the most efficient comparison of the different treatments?
4. What are the valid statistical procedures for making the comparisons?
5. What types of information should be included in a final report to document the age groups for which the laser treatment was most effective?

Collecting the Data

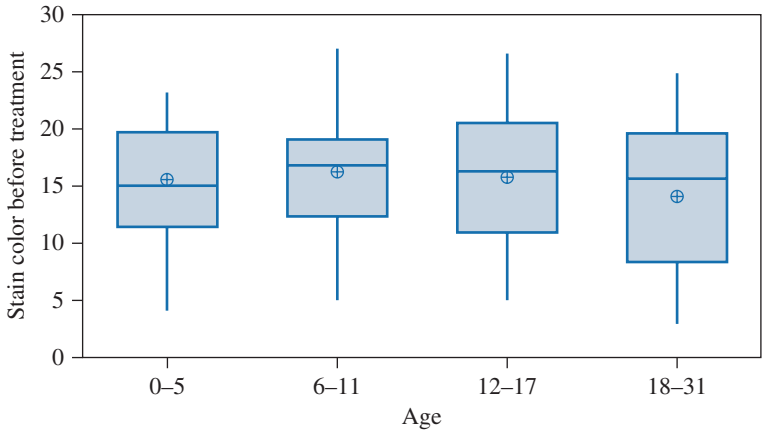
One hundred patients, 31 years of age or younger, with a previously untreated port-wine stain were selected for inclusion in the study. During the first consultation, the extent and location of the port-wine stain were recorded. Four age groups of 25 patients each were determined for evaluating whether the laser treatment was more effective for younger patients. Enrollment in an age group ended as soon as 25 consecutive patients had entered the group. A series of treatments was required to achieve optimal clearance of the stain. Before the first treatment, color slides were taken of each patient by a professional photographer in a studio under standardized conditions. The color of the skin was measured using a chromometer. The reproducibility of the color measurements was analyzed by measuring the same location twice in a single session before treatment. For each patient, subsequent color measurements were made at the same location. Treatment was discontinued if either the port-wine stain had disappeared or the three previous treatments had not resulted in any further lightening of the stain. The outcome measure of each patient was the reduction in the difference in color between the skin with the port-wine stain and the contralateral healthy skin.

Eleven of the 100 patients were not included in the final analysis due to a variety of circumstances that occurred during the study period. A variety of baseline characteristics was recorded for the 89 patients: the sex of the patient, the surface area and location of the port-wine stain, and any additional medical conditions that might have had implications for the effectiveness of the treatment. Researchers also recorded treatment characteristics such as the average number of visits, level of radiation exposure, number of laser pulses per visit, and occurrence of headaches after treatment. For all variables, there were no significance differences among the four age groups with respect to these characteristics.

Summarizing the Data

The two main variables of interest to the researchers were the difference in color between the port-wine stain and contralateral healthy skin before treatment and the improvement in this difference in color after a series of treatments. The before-treatment differences in color are presented in Figure 8.10. The boxplots demonstrate there were not sizable differences in color among the four groups. This is important, since if the groups differed prior to treatment, then the effect

FIGURE 8.10
Boxplots of stain color by age group (means are indicated by circles)



of age group on the effectiveness of the treatment may have been masked by the preexisting differences.

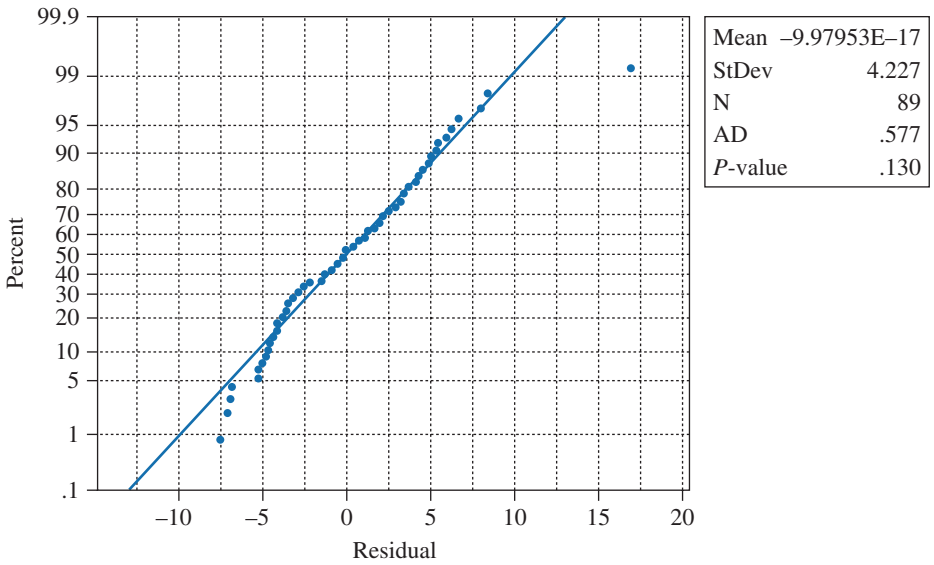
The improvement after treatment in the differences in color between the stain and healthy skin for each of the patients is given in Table 8.25. (These values were simulated using the summary statistics given in the original paper.)

The summary statistics for the above data were provided in Table 8.3. Boxplots of the improvement in stain color for the four age groups are displayed in Figure 8.2.

TABLE 8.25
Improvement in color of port-wine stains by age group

Patient	0–5 Years	6–11 Years	12–17 Years	18–31 Years
1	9.6938	13.4081	10.9110	1.4352
2	7.0027	8.2520	10.3844	10.7740
3	10.3249	12.0098	6.4080	8.4292
4	2.7491	7.4514	13.5611	4.4898
5	.5637	6.9131	3.4523	13.6303
6	8.0739	5.6594	9.5427	4.1640
7	.1440	8.7352	10.4976	5.4684
8	8.4572	.2510	4.6775	4.8650
9	2.0162	8.9991	24.7156	3.0733
10	6.1097	6.6154	4.8656	12.3574
11	9.9310	6.8661	.5023	7.9067
12	9.3404	5.5808	7.3156	9.8787
13	1.1779	6.6772	10.7833	2.3238
14	1.3520	8.2279	9.7764	6.7331
15	.3795	.1883	3.6031	14.0360
16	6.9325	1.9060	9.5543	.6678
17	1.2866	7.7309	5.3193	2.7218
18	8.3438	7.9143	3.0053	2.3195
19	9.2469	1.8724	11.0496	1.6824
20	.7416	12.5082	2.8697	1.8150
21	1.1072	6.2382	.1082	5.9665
22		11.2425		.5041
23		6.8404		5.4484
24		11.2774		

FIGURE 8.11



Analyzing the Data

The objective of the research study was to evaluate whether the treatment of port-wine stains was more effective for younger than for older children. We observed in Figure 8.2 that two of the age groups had outliers but otherwise that the boxplots had boxes of nearly of the same width and had whiskers of generally the same length. The means and medians were of a similar size for each of the four age groups. Thus, the assumptions of AOV would appear to be satisfied. To confirm this observation, we computed the residuals and plotted them in a normal probability plot (see Figure 8.11). From this plot, we can observe that, with the exception of one data value, the points fall nearly on a straight line. Also, the *p*-value for the test of the null hypothesis that the data have a normal distribution is .130. Thus, there is a strong confirmation that the four populations of improvements in skin color have normal distributions.

Next, we can check on the equal variance assumption by using the BFL test. For the BFL test, we obtain

$$L = 1.05 \text{ with } p\text{-value} = 1 - pf(1.05, 3, 85) = .3748$$

This implies there is not significant evidence that the four population variances differ. Based on the data, there is not significant evidence that the normality and equal variance conditions of the AOV procedure are violated. The condition of independence of the data would be checked by discussing with the researchers the manner in which the study was conducted. The sequencing of treatment and the evaluation of the color of the stains should have been performed such that the determination of improvement in color of one patient would not in any manner affect the determination of improvement in color of any other patient. The kinds of problems that may arise in this type of experiment and that can cause dependencies in the data include equipment problems, technician biases, any relationships between patients, and other similar factors.

The research hypothesis is that the mean improvement in stain color after treatment is different for the four age groups:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ versus H_a: At least one of the means differs from the rest.

The computer output for the AOV table is given here.

```
One-way ANOVA: 0-5 Years, 6-11 Years, 12-17 Years, 18-31 Years

Source  DF      SS    MS     F      P
Factor   3   108.0  36.0  1.95  0.128
Error   85  1572.5  18.5
Total   88  1680.5

S = 4.301   R-Sq = 6.43%  R-Sq(adj) = 3.12%


                                  Individual 95% CIs For Mean Based on
                                  Pooled StDev
Level          N    Mean   StDev ----+---------+---------+---------+-----
0-5 Years     21   4.999   3.916 (---------*---------)
6-11 Years    24   7.224   3.564           (---------*---------)
12-17 Years   21   7.757   5.456             (---------*---------)
18-31 Years   23   5.682   4.147    (---------*---------)
                                  ----+---------+---------+---------+-----

                                     4.0       6.0       8.0      10.0

Pooled StDev = 4.301
```

From the output, the p-value for the F test is .128. Thus, there is not a significant difference in the mean improvements for the four groups. We can also compute 95% confidence intervals for the mean improvements. The four intervals are provided in the computer output. They are computed using the pooled standard deviation, $\hat{\sigma} = \sqrt{MSW} = \sqrt{19.7} = 4.44$ with df = 85. Thus, the intervals are of the form

$$\bar{y}_{i.} \pm t_{.025,85}\,\hat{\sigma}/\sqrt{n_i} = \bar{y}_{i.} \pm (1.99)(4.44)/\sqrt{n_i}$$

The four intervals are presented in Table 8.26.

TABLE 8.26
Confidence intervals for age groups

Age Group	$\bar{y}_{i.}$	95% C.I.
0–5	4.999	(3.07, 6.93)
6–11	7.224	(5.42, 9.03)
12–7	7.760	(5.83, 9.69)
18–31	5.682	(3.84, 7.52)

From these confidence intervals, we can compare the mean improvements in stain color for the four groups. The youngest age group has the smallest improvement, but its upper bound is greater than the lower bound for the age group having the greatest improvement. The problem with this type of decision making is that the confidence intervals are not simultaneous confidence intervals, and, hence, we cannot attribute a level of certainty to our conclusions. In the next chapter, we will present simultaneous confidence intervals for the difference in treatment means and hence will be able to decide which pairs of treatments in fact are significantly

different. However, in our research study, we can safely conclude that all pairs of treatment means are not significantly different, since the AOV F test failed to reject the null hypothesis.

The researchers did not confirm the hypothesis that treatment of port-wine stains at an early age is more effective than treatment at a later age. The researchers did conclude that their results had implications for the timing of therapy in children. Although facial port-wine stains can be treated effectively and safely early in life, treatment at a later age leads to similar results. Therefore, the age at which therapy is initiated should be based on a careful weighing of the anticipated benefit and the discomfort of treatment.

Reporting the Conclusions

We would need to write a report summarizing our findings of this prospective study of the treatment of port-wine stains. We would need to include the following:

1. Statement of objective for study
2. Description of study design and data collection procedures
3. Discussion of why the results from 11 of the 100 patients were not included in the data analysis
4. Numerical and graphical summaries of data sets
5. Description of all inference methodologies
   - AOV table and F test
   - t-based confidence intervals on means
   - Verification that all necessary conditions for using inference techniques were satisfied
6. Discussion of results and conclusions
7. Interpretation of findings relative to previous studies
8. Recommendations for future studies
9. Listing of data sets

8.8 Summary and Key Formulas

In this chapter, we presented methods for extending the results of Chapter 6 to include a comparison among t population means. An independent random sample is drawn from each of the t populations. A measure of the within-sample variability is computed as $s_W^2 = \text{SSW}/(n_T - t)$. Similarly, a measure of the between-sample variability is obtained as $s_B^2 = \text{SSB}/(t - 1)$.

The decision to accept or reject the null hypothesis of equality of the t population means depends on the computed value of $F = s_B^2/s_W^2$. Under H_0, both s_B^2 and s_W^2 estimate σ^2, the variance common to all t populations. In Chapter 14, it will be shown that under the alternative hypothesis, s_B^2 estimates $\sigma^2 + \theta$, where θ is a positive quantity, whereas s_W^2 still estimates σ^2. Thus, large values of F indicate a rejection of H_0. Critical values for F are obtained from Table 8 in the Appendix for $\text{df}_1 = t - 1$ and $\text{df}_2 = n_T - t$. This test procedure, called an analysis of variance, is usually summarized in an analysis of variance (AOV) table.

You might be puzzled at this point by the following: Suppose we reject H_0 and conclude that at least one of the means differs from the rest. Which ones differ from the others? This chapter has not answered this question; Chapter 9 attacks this problem through procedures based on multiple comparisons.

In this chapter, we also discussed the assumptions underlying an analysis of variance for a completely randomized design. Independent random samples are absolutely necessary. The assumption of normality is least critical because we are dealing with means and the Central Limit Theorem holds for reasonable sample sizes. The equal variance assumption is critical only when the sample sizes are markedly different; this is a good argument for equal (or nearly equal) sample sizes. A test for equality of variances makes use of the BFL test.

Sometimes the sample data indicate that the population variances are different. Then, when the relationship between the population mean and the population standard deviation is either known or suspected, it is convenient to transform the sample measurements y to new values y_T to stabilize the population variances, using the transformations suggested in Table 8.15. These transformations include the square root, logarithmic, arcsin, and many others.

The topics in this chapter are certainly not covered in exhaustive detail. However, the material is sufficient for training the beginning researcher to be aware of the assumptions underlying his or her project and to consider either running an alternative analysis (such as using a nonparametric statistical method, the Kruskal–Wallis test) when appropriate or applying a transformation to the sample data.

Key Formulas

1. Analysis of variance for a completely randomized design

$$\text{SSB} = \sum_i n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$\text{SSW} = \sum_{i,j}(y_{ij} - \bar{y}_{i.})^2 = \sum_i (n_i - 1)s_i^2$$

$$\text{TSS} = \sum_{i,j}(y_{ij} - \bar{y}_{..})^2 = \text{SSB} + \text{SSW}$$

2. Model for a completely randomized design

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where $\mu_i = \mu + \tau_i$

3. Conditions imposed on model:
   a. The t populations have normal distributions
   b. $\sigma_1^2 = \cdots = \sigma_t^2 = \sigma^2$
   c. Data consist of t independent random samples

4. Check whether conditions are satisfied:
   a. Normality: Plots of residuals, $e_{ij} = y_{ij} - \bar{y}_{i.}$
   b. Homogeneity of variance: BFL test
   c. Independence: Careful review of how experiment or study was conducted

5. $100(1 - \alpha)\%$ confidence intervals for population means μ_i

$$\bar{y}_{i.} \pm t_{\alpha/2,\, n_T - t}\frac{\hat{\sigma}}{\sqrt{n_i}}$$

where $\hat{\sigma} = \sqrt{\text{MSW}}$

6. Kruskal–Wallis test (when population distributions are very nonnormal)

   H_0: The k population distributions are identical.
   H_a: The k population distributions are shifted from each other.

$$\text{T.S.} = \frac{12}{n_T(n_T + 1)}\sum_{i=1}^{k}\frac{T_i^2}{n_i} - 3(n_T + 1)$$

8.9 Exercises

8.1 Introduction

Med. **8.1** For the port-wine stains research study, answer the following:

- **a.** What are the populations of interest?
- **b.** What are some factors besides change in skin color that may be of interest to the investigators?

Med. **8.2** For the port-wine stains research study, do the following:

- **a.** Describe how the subjects in this experiment could have been selected so as to satisfy the randomization requirements.
- **b.** State several research hypotheses that may have been of interest to the researchers.

8.2 A Statistical Test About More Than Two Population Means: An Analysis of Variance

Theory **8.3** A number of new techniques for teaching reading have been proposed in the educational literature. A researcher designs the following study to evaluate three of these new methods along with the standard method, which has been used for a number of years. In a large school district, five elementary schools are selected for inclusion in the study. Four third-grade teachers are randomly selected in each of the five schools, and the four reading techniques are randomly assigned to the teachers. The teachers participate in a 2-week workshop to learn the fine points of their assigned technique. The students in the 20 classrooms are given a standardized reading examination at the end of the semester, with the average score in each classroom used as the measure of the effectiveness of the teaching technique. Thus, there are five measurements of the effectiveness of each of the four teaching techniques.

- **a.** What are the populations of interest in this study?
- **b.** The conclusions of this study can properly be inferred for what populations?
- **c.** Would it be appropriate to use the AOV F test to evaluate whether there is a difference in the average scores of the four teaching techniques?

Theory **8.4** In Example 8.3, suppose the organization wanted to compare the mean test scores of Catholic priests and Methodist ministers. Note that it appears based on the data that these two groups have the same variance. What is the gain in using a two-sample t-test having s_W^2 in the denominator as opposed to using the conventional pooled t test with s_p^2, the average of the sample variances for the Catholic priests and Methodist ministers?

Theory **8.5** Consider an experiment designed to compare four treatment means—μ_1, μ_2, μ_3, and μ_4—using sample sizes of size n_1, n_2, n_3, and n_4 and sample variances s_1^2, s_2^2, s_3^2, and s_4^2.

- **a.** Suppose the sample sizes are the same: $n_1 = n_2 = n_3 = n_4$. Show that s_W^2 is the average of the four sample variances: $s_W^2 = (s_1^2 + s_2^2 + s_3^2 + s_4^2)/4$.
- **b.** Does this hold if the sample sizes are not equal? If not, why not just use the average?

Ag. **8.6** A large laboratory has four types of devices used to determine the pH of soil samples. The laboratory wants to determine whether there are differences in the average readings given by these devices. The lab uses 24 soil samples having known pH in the study and randomly assigns six of the samples to each device. The soil samples are tested, and the response recorded for each sample is the difference between the pH reading of the device and the known pH of the soil. These values, along with summary statistics, are given in the following table.

	Sample					
Device	**1**	**2**	**3**	**4**	**5**	**6**
A	−.307	−.294	.079	.019	−.136	−.324
B	−.176	.125	−.013	.082	.091	.459
C	.137	−.063	.240	−.050	.318	.154
D	−.042	.690	.201	.166	.219	.407

a. Based on your intuition, is there evidence to indicate any difference among the mean differences in pH readings for the four devices?
b. Run an analysis of variance to confirm or reject your conclusion in part (a). Use $\alpha = .05$.
c. Compute the p-value of the F test in part (b).
d. What conditions must be satisfied for your analysis in parts (b) and (c) to be valid?
e. Suppose the 24 soil samples have widely different pH values. What problems may occur by simply randomly assigning the soil samples to the different devices?

Ag. **8.7** It is conjectured that when fields are overgrazed by cattle there will be a substantial reduction in the available grass during the subsequent grazing season due to the compaction of the soil. A horticulturist at the state agricultural experiment station designs a study to evaluate the conjecture. Twenty-one plots of land of nearly the same soil texture and suitable for grazing are selected for the study. Three grazing regimens selected for evaluation are randomly assigned to 7 plots each. After the 21 plots are subjected to the grazing regimens for four months, the researcher randomly selects 10 soil cores from each plot and measures the bulk density (g/cm^3) in each soil core. The mean soil density of the 10 cores from each plot is given in the following table.

Grazing Regimen	**Soil Density (g/cm^3)**						
Continuous grazing	2.05	3.05	3.12	1.59	3.83	1.53	1.44
Three-week grazing, one-week no grazing	1.20	1.48	3.54	1.03	1.45	1.40	2.68
Two-week grazing, two-week no grazing	1.23	1.66	1. 70	1.29	1.26	1.05	2.35

a. Do the grazing regimens appear to yield different degrees of effect on the amount of compacting in the soil? Justify your answer using an $\alpha = .05$ test.
b. Provide the level of significance of your test.
c. Do any of the conditions necessary for conducting your test appear to be violated? Justify your answer.

8.3 The Model for Observations in a Completely Randomized Design

Theory **8.8** An experiment is designed to compare the means of four populations. Suppose the population means are given as follows:

$$\mu_1 = 18 \quad \mu_2 = 28 \quad \mu_3 = 7 \quad \mu_4 = 31$$

Using the relationship $\mu_i = \mu + \tau_i$ with the constraint $\tau_4 = 0$, compute the values of μ, τ_1, τ_2, τ_3, and τ_4.

Con. **8.9** Refer to Example 8.1.
a. For the model $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, what are the values of t, n_1, n_2, and n_3?
b. Using the observed data, provide estimates of μ, τ_1, τ_2, τ_3, and σ without the constraint $\tau_3 = 0$.
c. Using the observed data, provide estimates of μ, τ_1, τ_2, τ_3, and σ with the constraint $\tau_3 = 0$.
d. Compare the differences in the two sets of estimates produced in parts (b) and (c). This illustrates the importance of knowing what constraints are imposed by software programs when estimates are contained in the output of an analysis.

Con. **8.10** Refer to Example 8.4.
a. For the model $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, what are the values of t, n_1, n_2, n_3, and n_4?
b. Using the observed data, provide estimates of μ, τ_1, τ_2, τ_3, and σ with the constraint $\tau_4 = 0$.

Theory **8.11** In a study of five populations with five equal sample sizes of $n_i = 20$, the 100 data values produced a mean square within samples of $s_W^2 = 0$. Without having access to the 100 data values, answer the following questions about the sample means and the residuals?
a. Are the five sample means equal?
b. What can be concluded about the 100 sample residuals: $e_{ij} = y_{ij} - \bar{y}_i$?

Con. **8.12** Refer to Example 8.1.
- **a.** Using the observed data, compute the 15 sample residuals: $e_{ij} = y_{ij} - \bar{y}_{i\cdot}$.
- **b.** Using the 15 residuals, verify that $s_W^2 = \frac{1}{12}\sum_{ij} e_{ij}^2$.
- **c.** Do the data indicate any violations in the conditions for conducting the AOV F test?

Med. **8.13** Refer to Example 8.2.
- **a.** Demonstrate that s_W^2 is not the average of the three sample variances: s_1^2, s_2^2, and s_3^2.
- **b.** Do the data indicate any violations in the conditions for conducting the AOV F test?

Med. **8.14** Refer to Example 8.6.
- **a.** Do the data indicate that the populations are not normally distributed?
- **b.** Do the transformed data appear to have a normal distribution?
- **c.** Does a transformation that produces data having equal variances guarantee that the transformed data will be normally distributed?

Med. **8.15** Refer to Example 8.3.
- **a.** Do the data indicate that the populations are not normally distributed?
- **b.** Find a transformation of the data such that the transformed data appears to have a normal distribution.

8.5 An Alternative Analysis: Transformations of the Data

Env. **8.16** Refer to Example 8.4.
- **a.** Apply the AOV F test to the original measurements using $\alpha = .05$.
- **b.** Apply the AOV F test to the transformed data using $\alpha = .05$.
- **c.** Did transforming the data alter your conclusion as to whether the oxygen content is related to the distance to the mouth of the Mississippi River?

Pol. Sci. **8.17** Refer to Example 8.6.
- **a.** Apply the AOV F test to the original measurements using $\alpha = .05$.
- **b.** Apply the AOV F test to the transformed data using $\alpha = .05$.
- **c.** Did transforming the data alter your conclusion as to whether there is a difference in the four geographical regions with respect to their opinion of the EPA regulations on air pollution?

Engin. **8.18** Refer to Example 7.8. The consumer testing agency was interested in evaluating whether there was a difference in the mean percentage increases in mpg of the three additives. In Example 7.9, we showed that the data did not appear to have a normal distribution.
- **a.** Apply the natural logarithm transformation to the data. Do the conditions for applying the AOV F test appear to hold for the transformed data?
- **b.** Test for a difference in the means of the three additives using $\alpha = .05$.

Bio. **8.19** Refer to Exercise 7.18.
- **a.** The biologist hypothesized that the mean weight of deer raised in a zoo would differ from the mean weight of deer raised either in the wild or on a ranch. Do the conditions necessary for applying the AOV F test appear to be valid?
- **b.** If the conditions for the AOV F test are satisfied, then conduct the test to evaluate the biologist's claim. If not, then suggest a transformation, and conduct the test on the transformed data.

Edu. **8.20** The use of computers as an instructional aid is widely advocated as a means to capture the attention of the current computer literate generation of students. A study was designed to assess the effectiveness of using computers as a supplement to the standard mode of instruction. Forty students in an alternative school were randomly assigned to one of four methods of teaching basic math skills. The four methods were lectures only (L), lectures with remedial text book assistance (L/R), lectures with computer assistance (L/C), and computer instruction only (C). After a 10-week instructional period, an exam evaluating basic math skills was taken by the students. The difference in the scores on this exam and on an exam given just prior to the 10-week instructional period for each student is given in the following table. A few of the students did not complete the program, thus producing an unequal number of students in the four modes of instruction.

The researchers want to determine which method of instruction produces the largest increase in test scores.

	Student									
Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
L	9	2	2	6	16	11	9	0	4	2
L/R	5	2	3	11	16	11	3			
L/C	9	12	2	17	12	20	20	31	21	
C	17	12	26	1	47	27	−8	10	20	

a. Which method of instruction appears to produce the largest gain in scores?
b. Is there significant evidence of a difference in the mean gains for the four methods of instruction?
c. Do the conditions for conducting statistical tests appear to be satisfied? Justify your conclusions with graphs/tests.
d. What is the target population for this study?
e. Do the data collected in this study allow inferences to the target population?

Soc. **8.21** Refer to Exercise 3.55.

a. The state legislative committee in charge of allocations for food stamps wanted to determine if there was a difference in the mean food expenditures among the five family sizes. Do the conditions necessary for applying the AOV F test appear to be valid?
b. If the conditions for the AOV F test are satisfied, then conduct the test to evaluate whether there is a difference in the five food expenditure means. If not, then suggest a transformation, and conduct the test on the transformed data.

8.22 Refer to Example 8.5. In many situations in which the difference in variances is not too great, the results from the AOV comparisons of the population means of the transformed data are very similar to the results that would have been obtained using the original data. In these situations, the researcher is inclined to ignore the transformations because the scale of the transformed data is not relevant to the researcher. Thus, confidence intervals constructed for the means using the transformed data may not be very relevant. One possible remedy for this problem is to construct confidence intervals using the transformed data and then perform an inverse transformation of the endpoints of the intervals. Then we would obtain a confidence interval with values having the same units of measurement as the original data.

a. Test the hypothesis that the mean hours of relief for patients from the three treatments differs using $\alpha = .05$. Use the original data.
b. Place 95% confidence intervals on the mean hours of relief for the three treatments.
c. Repeat the analysis in parts (a) and (b) using the transformed data.
d. Comment on any differences in the results of the test of hypotheses.
e. Perform an inverse transformation on the endpoints of the intervals constructed in part (c). Compare these intervals to the ones constructed in part (b).

8.6 A Nonparametric Alternative: The Kruskal–Wallis Test

Engin. **8.23** In a 1996 article published in **Technometrics, *(Martz, Kvan, Abramson, 1996)***, the authors discuss the reliability of nuclear-power-plant emergency generators. To control the risk of damage to the nuclear core during accidents at nuclear plants, the reliability of emergency diesel generators (EDGs) to start on demand must be maintained at a very high level. At each nuclear power plant, there are a number of such generators. An overall measure of reliability is obtained by counting the number of times the EDGs successfully work when needed. The table here provides the number of successful demands for implementation of an EDG between each subsequent failure in an EDG for all the EDGs at each of seven nuclear power plants. A regulatory agency wants to determine if there is a difference in the reliabilities of the seven nuclear power plants.

Plant	n_i	Number of Times EDG Works																
A	34	28	50	193	55	4	7	174	76	10	0	10	84	0	9	1	0	62
		26	15	226	54	46	128	4	105	40	4	273	164	7	55	41	26	6
B	15	2	11	75	6	1	12	4	6	64	3	0	3	1	20	78		
C	17	142	110	3	273	54	32	3	40	23	30	17	7	12	6	12	7	5
D	8	64	29	1	3	8	29	4	60									
E	12	139	21	214	67	174	1	9	2	119	237	110	71					
F	7	18	108	9	8	17	88	28										
G	10	0	6	0	16	1	58	13	36	33	19							

a. Do the conditions necessary for conducting the AOV F test appear to be satisfied by these data?
b. Because the data are counts of the number of successes for the EDGs, the Poisson model may be an alternative to the normal-based analysis. Apply a transformation to the data, and then apply the AOV F test to the transformed data.
c. As a second alternative analysis that has fewer restrictions, answer the agency's question by applying the Kruskal–Wallis test to the reliability data.
d. Compare your conclusions to parts (a)–(c). Which of the three procedures provides the conclusion about which you feel most confident?

Env. **8.24** Refer to Example 8.4.
a. Apply the Kruskal–Wallis test to determine if there is a difference in the distributions of oxygen content for the various distances to the mouth of the Mississippi River.
b. Does your conclusion differ from the conclusion reached in Exercise 8.16?

Med. **8.25** Refer to Example 8.5.
a. Apply the Kruskal–Wallis test to determine if there is a difference in the distributions of pain reduction for the three analgesics.
b. Does your conclusion differ from the conclusion reached in Exercise 8.22?

Med. **8.26** Refer to Example 8.6.
a. Apply the Kruskal–Wallis test to determine if there is a difference in the distributions of opinions across the four geographical regions.
b. Does your conclusion differ from the conclusion reached in Exercise 8.17?

Engin. **8.27** *Wludyka and Nelson (1997)* describe the following study. In the manufacture of soft contact lenses, a monomer is injected into a plastic frame, the monomer is subjected to ultraviolet light and heated (the time, temperate, and light intensity are varied), the frame is removed, and the lens is hydrated. It is thought that temperature can be manipulated to target the power (strength of the lens), so comparing the variability in power is of interest. The data are coded deviations from the target power using monomers from five different suppliers given below.

	Sample								
Supplier	1	2	3	4	5	6	7	8	9
1	191.9	189.1	190.9	183.8	185.5	190.9	192.8	188.4	189.0
2	178.2	174.1	170.3	171.6	171.7	174.7	176.0	176.6	172.8
3	156.6	158.4	157.7	154.1	152.3	161.5	158.1	150.9	156.9
4	125.8	132.4	132.2	133.0	133.2	125.9	132.9	142.6	135.5
5	218.6	208.4	187.1	199.5	202.0	211.1	197.6	204.4	206.8

a. Do the suppliers appear to differ in their levels of variability? Use $\alpha = .05$.
b. Is there significant evidence of a difference in the mean deviations for the five suppliers? Use an $\alpha = .05$ AOV F test.

c. Apply the Kruskal–Wallis test to evaluate differences in the distributions of the deviations for the five suppliers? Use $\alpha = .05$.

d. Suppose a difference in mean deviations of 20 units would have commercial consequences for the manufacturer of the lenses. Does there appear to be a *practical* difference in the materials from the five suppliers?

Ag. **8.28** The Agricultural Experiment Station of a university tested two different herbicides and their effects on crop yield. From 90 acres set aside for the experiment, the station used herbicide 1 on a random sample of 30 acres and herbicide 2 on a second random sample of 30 acres; they used the remaining 30 acres as a control. At the end of the growing season, the yields (in bushels per acre) were as follows:

Herbicide 1	81.2	81.1	79.9	84.6	80.4	74.4	81.7	90.1	102.4	89.2
	92.0	91.7	75.9	95.1	76.1	83.0	88.0	80.5	73.6	80.4
	103.2	85.9	73.6	80.0	82.4	79.5	99.8	96.6	81.3	94.7
Herbicide 2	94.8	90.9	85.2	83.3	95.5	85.4	87.1	89.6	83.7	88.7
	91.4	85.4	89.0	92.4	85.0	91.0	89.2	100.9	88.5	90.3
	87.6	80.7	90.0	101.0	92.1	97.9	92.5	88.8	89.4	100.1
Control	94.7	79.5	91.4	82.6	96.9	85.4	80.8	90.6	88.6	80.8
	78.1	82.5	93.5	83.1	90.5	89.2	82.0	84.1	90.1	84.5
	81.2	92.4	90.5	82.0	106.6	96.9	76.1	101.8	77.5	88.8

a. Use these data to conduct a one-way analysis of variance to test whether there is a difference in the mean yields. Use $\alpha = .05$.

b. Construct 95% confidence intervals on the mean yields μ_i.

c. Which of the mean yields appear to be different from the control?

Hort. **8.29** Researchers from the Department of Fruit Crops at a university compared four different preservatives to be used in freezing strawberries. The researchers prepared the yield from a strawberry patch for freezing and randomly divided it into four equal groups. Within each group, they treated the strawberries with the appropriate preservative and packaged them into eight small plastic bags for freezing at 0°C. The bags in group I served as a control group, while those in groups II, III, and IV were assigned one of three newly developed preservatives. After all 32 bags of strawberries were prepared, they were stored at 0°C for a period of 6 months. At the end of this time, the contents of each bag were allowed to thaw and then rated on a scale of 1 to 10 points for discoloration. (Note that a low score indicates little discoloration.) The ratings are given here:

Group I	10	8	7.5	8	9.5	9	7.5	7
Group II	6	7.5	8	7	6.5	6	5	5.5
Group III	3	5.5	4	4.5	3	3.5	4	4.5
Group IV	2	1	2.5	3	4	3.5	2	2

a. Assess whether the conditions needed to use AOV techniques are satisfied with this data set.

b. Test whether there is a difference in the mean ratings using $\alpha = .05$.

c. Place a 95% confidence interval on the mean rating for each of the groups.

8.30 Refer to Exercise 8.29. In many situations in which the response is a rating rather than an actual measurement, it is recommended that the Kruskal–Wallis test be used.

a. Apply the Kruskal–Wallis test to determine whether there is a shift in the distribution of ratings for the four groups.

b. Is the conclusion reached using the Kruskal–Wallis test consistent with the conclusion reached in Exercise 8.29 using AOV?

H.R. **8.31** Salary disputes and their eventual resolutions often leave both employers and employees embittered by the entire ordeal. To assess employee reactions to a recently devised salary and fringe benefits plan, the personnel department obtained random samples of 15 employees from each of three divisions in the company: manufacturing, marketing, and research. The personnel staff asked each employee sampled to respond (in confidence) to a series of questions. Several employees refused to cooperate, as reflected in the unequal sample sizes. The data are given here:

Manufacturing	18.79	22.46	31.99	24.74	29.52	20.25	31.64
	28.66	27.97	28.19	20.22	29.18		
Marketing	27.63	31.22	35.33	31.06	36.50	29.92	33.18
	37.03	35.22	37.89	23.01	37.81	33.41	29.361
Research	26.64	28.90	32.05	26.54	27.12	35.78	26.28
	31.90	25.70	25.44	33.41			

The data given above are the average responses from the employees, with larger scores reflecting a higher degree of satisfaction with management.

a. Write a model for this situation. Make sure to identify all the terms in your model.
b. Based on the summary of the scored responses, is there significant evidence of a difference among the three divisions with respect to their levels of satisfaction with management?

Ag. **8.32** Researchers record the yields of corn, in bushels per plot, for four different varieties of corn, A, B, C, and D. In a controlled greenhouse experiment, the researchers randomly assign each variety to 8 of 32 plots available for the study. The yields are listed here:

A	2.5	3.6	2.8	2.7	3.1	3.4	2.9	3.5
B	3.6	3.9	4.1	4.3	2.9	3.5	3.8	3.7
C	4.3	4.4	4.5	4.1	3.5	3.4	3.2	4.6
D	2.8	2.9	3.1	2.4	3.2	2.5	3.6	2.7

a. Write an appropriate statistical model.
b. Perform an analysis of variance on these data, and draw your conclusions. Use $\alpha = .05$.

8.33 Refer to Exercise 8.32. Perform a Kruskal–Wallis test (with $\alpha = .05$), and compare your results to those in Exercise 8.32.

Edu. **8.34** Doing homework is a nightly routine for most school-age children. The article ***"Family Involvement with Middle-Grades Homework: Effects of Differential Prompting" (Balli, S. J., J. F. Wedman, and D. H. Demo, 1997)***, examines the question of whether parents' involvement with their children's homework is associated with improved academic performance. Seventy-four sixth graders and their families participated in the study. The students, similar in academic ability and background, were enrolled in one of three mathematics classes taught by the same teacher; researchers randomly assigned each class to one of the three treatment groups.

Group I, student/family prompt: Students were prompted to seek assistance from a family member, and family members were encouraged to provide assistance to the students.
Group II, student prompt: Students were prompted to seek assistance from a family member, but there was no specific encouragement of family members to provide assistance to the students.

Group III, no prompts: Students were not prompted to seek assistance from a family member nor were family members encouraged to provide assistance to the students.

The researchers gave the students a posttest, with the results given here:

Treatment Group	Number of Students	Mean Posttest Score
Student/family prompt	22	68%
Student prompt	22	66%
No prompt	25	67%

The researchers concluded that higher levels of family involvement were not associated with higher student achievement in this study.

a. What is the population of interest in this study?
b. Based on the data collected, to what population can the results of this study be attributed?
c. What is the effective sample for each of the treatment groups; that is, how many experimental units were randomly assigned to each of the treatment groups?
d. What criticisms would you have for the design of this study?
e. Suggest an improved design for addressing the research hypothesis that family involvement improves student performance in mathematics classes.

Gov.

8.35 In a 1994 Senate subcommittee hearing, an executive of a major tobacco company testified that the accusation that nicotine was added to cigarettes was false. Tobacco company scientists stated that the amount of nicotine in cigarettes was completely determined by the size of the tobacco leaf, with smaller leaves having greater nicotine content. Thus, the variation in nicotine content in cigarettes occurred due to a variation in the size of the tobacco leaves and was not due to any additives placed in the cigarettes by the company. Furthermore, the company argued that the size of the leaves varied depending on the weather conditions during the growing season, over which they had no control. To study whether smaller tobacco leaves had a higher nicotine content, a consumer health organization conducted the following experiment. The major factors controlling leaf size are the temperature and the amount of water received by the plants during the growing season. The experimenters created four types of growing conditions for tobacco plants. Condition A was average temperature and rainfall amounts. Condition B was lower than average temperature and rainfall conditions. Condition C was higher than average temperature with lower than average rainfall. Finally, condition D was higher than average temperature and rainfall. The scientists then planted 10 tobacco plants under each of the four conditions in a greenhouse where temperature and amount of moisture were carefully controlled. After growing the plants, the scientists recorded the leaf size and nicotine content, which are given here:

Plant	A Leaf Size	B Leaf Size	C Leaf Size	D Leaf Size
1	27.7619	4.2460	15.5070	33.0101
2	27.8523	14.1577	5.0473	44.9680
3	21.3495	7.0279	18.3020	34.2074
4	31.9616	7.0698	16.0436	28.9766
5	19.4623	0.8091	10.2601	42.9229
6	12.2804	13.9385	19.0571	36.6827
7	21.0508	11.0130	17.1826	32.7229
8	19.5074	10.9680	16.6510	34.5668
9	26.2808	6.9112	18.8472	28.7695
10	26.1466	9.6041	12.4234	36.6952

Plant	A Nicotine	B Nicotine	C Nicotine	D Nicotine
1	10.0655	8.5977	6.7865	9.9553
2	9.4712	8.1299	10.9249	5.8495
3	9.1246	11.3401	11.3878	10.3005
4	11.3652	9.3470	9.7022	9.7140
5	11.3976	9.3049	8.0371	10.7543
6	11.2936	10.0193	10.7187	8.0262
7	10.6805	9.5843	11.2352	13.1326
8	8.1280	6.4603	7.7079	11.8559
9	10.5066	8.2589	7.5653	11.3345
10	10.6579	5.0106	9.0922	10.4763

a. Perform a one-way analysis of variance to test whether there is a significant difference in the average leaf sizes under the four growing conditions. Use $\alpha = .05$.
b. What conclusions can you reach concerning the effect of growing conditions on the average leaf size?
c. Perform a one-way analysis of variance to test whether there is a significant difference in the average nicotine contents under the four growing conditions. Use $\alpha = .05$.
d. What conclusions can you reach concerning the effect of growing conditions on the average nicotine content?
e. Based on the conclusions you reached in parts (b) and (d), do you think the testimony of the tobacco companies' scientists is supported by this experiment? Justify your conclusions.

8.36 Do the nicotine content data in Exercise 8.35 suggest violations of the AOV conditions? If you determine that the conditions are not met, perform an alternative analysis, and compare your results to those of Exercise 8.35.

Ag.

8.37 Scientists conducted an experiment to test the effects of five different diets on turkeys. They randomly assigned six turkeys to each of the five diet groups and fed them for a fixed period of time.

Group	Weight Gained (pounds)
Control diet	4.1, 3.3, 3.1, 4.2, 3.6, 4.4
Control diet + level 1 of additive A	5.2, 4.8, 4.5, 6.8, 5.5, 6.2
Control diet + level 2 of additive A	6.3, 6.5, 7.2, 7.4, 7.8, 6.7
Control diet + level 1 of additive B	6.5, 6.8, 7.3, 7.5, 6.9, 7.0
Control diet + level 2 of additive B	9.5, 9.6, 9.2, 9.1, 9.8, 9.1

a. Plot the data separately for each sample.
b. Compute $\bar{y}$ and s^2 for each sample.
c. Is there any evidence of unequal variances or nonnormality? Explain.
d. Assuming that the five groups were comparable with respect to initial weights of the turkeys, use the weight-gained data to draw conclusions concerning the different diets. Use $\alpha = .05$.

8.38 Run a Kruskal–Wallis test for the data of Exercise 8.37. Do these results confirm what you concluded from an analysis of variance? What overall conclusions can be drawn? Use $\alpha = .05$.

Hort.

8.39 Some researchers have conjectured that stem-pitting disease in peach tree seedlings might be related to the presence or absence of nematodes in the soil. Hence, weed and soil treatment using herbicides might be effective in promoting seedling growth. Researchers conducted an experiment to compare peach tree seedling growth with soil and weeds using with one of three treatments:

A: Control (no herbicide)
B: Herbicide with Nemagone
C: Herbicide without Nemagone

The researchers randomly assigned 6 of the 18 seedlings chosen for the study to each treatment group. They treated soil and weeds in the growing areas for the three groups with the appropriate herbicide. At the end of the study period, they recorded the height (in centimeters) for each seedling. Use the following sample data to run an analysis of variance for detecting differences among the seedling heights for the three groups. Use $\alpha = .05$. Draw your conclusions.

Herbicide A	66	67	74	73	75	64
Herbicide B	85	84	76	82	79	86
Herbicide C	91	93	88	87	90	86

8.40 Refer to the data of Exercise 8.37. To illustrate the effect that an extreme value can have on conclusions from an analysis of variance, suppose that the weight gained by the fifth turkey in the level 2, additive B group was 15.8 rather than 9.8.

a. What effect does this have on the assumptions for an analysis of variance?

b. With 9.8 replaced by 15.8, if someone unknowingly ran an analysis of variance, what conclusions would he or she draw?

8.41 Refer to Exercise 8.40. What happens to the Kruskal–Wallis test if you replace the value 9.8 by 15.8? Might there be a reason to run both a Kruskal–Wallis test and an analysis of variance? Justify your answer.

Engin. **8.42** A small corporation makes insulation shields for electrical wires using three different types of machines. The corporation wants to evaluate the variation in the inside diameter dimensions of the shields produced by the machines. A quality engineer at the corporation randomly selects shields produced by each of the machines and records the inside diameter of each shield (in millimeters). She wants to determine whether the means and standard deviations of the three machines differ.

Shield	Machine A	Machine B	Machine C
1	18.1	8.7	29.7
2	2.4	56.8	18.7
3	2.7	4.4	16.5
4	7.5	8.3	63.7
5	11.0	5.8	18.9
6			107.2
7			19.7
8			93.4
9			21.6
10			17.8

a. Conduct a test for the homogeneity of the population variances. Use $\alpha = .05$.

b. Would it be appropriate to proceed with an analysis of variance based on the results of this test? Explain.

c. If the variances of the diameters are different, suggest a transformation that may alleviate their differences, and then conduct an analysis of variance to determine whether the mean diameters differ. Use $\alpha = .05$.

d. Compare the results of your analysis in part (c) to an analysis of variance on the original diameters.

e. How could the engineer have designed her experiment differently if she knew that the variances of machine B and machine C were so much larger than that of machine A?

8.43 The Kruskal–Wallis test is not as highly affected by unequal variances as the AOV test. Demonstrate this result by applying the Kruskal–Wallis test to both the original and the transformed data and comparing the conclusions reached in this analysis for the data of Exercise 8.42.

CHAPTER 9

Multiple Comparisons

9.1 Introduction and Abstract of Research Study

In Chapter 8, we introduced a procedure for testing the equality of t population means. We used the test statistic $F = s_B^2/s_W^2$ to determine whether the between-sample variability was large relative to the within-sample variability. If the computed value of F for the sample data exceeded the critical value obtained from Table 8 in the Appendix, we rejected the null hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$ in favor of the alternative hypothesis

H_a: At least one of the t population means differs from the rest.

Although rejection of the null hypothesis does give us some information concerning the population means, we do not know which means differ from each other. For example, does μ_1 differ from μ_2 or μ_3? Does μ_3 differ from the average of μ_2, μ_4, and μ_5? Is there an increasing trend in the treatment means $\mu_1, \ldots, \mu_t$? **Multiple-comparison procedures** and contrasts have been developed to answer questions such as these. Although many multiple-comparison procedures have been proposed, we will focus on just a few of the more commonly used methods. After studying these few procedures, you should be able to evaluate the results of most published material using multiple comparisons or to suggest an appropriate multiple-comparison procedure in an experimental situation.

multiple-comparison procedures

A word of caution: It is tempting to analyze only those comparisons that appear to be interesting after seeing the sample data. This practice has sometimes

data dredging
data snooping

been called **data dredging** or **data snooping,** and the confidence coefficient for a single comparison does not reflect the after-the-fact nature of the comparison. For example, we know from previous work that the interval estimate for the difference between two population means using the formula

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

has a confidence coefficient of $1 - \alpha$. Suppose we had run an analysis of variance to test the hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$$

for six populations but decided to compute a confidence interval for μ_1 and μ_2 only after we saw that the largest sample mean was $\bar{y}_1$. and the smallest was $\bar{y}_2$. In this situation, the confidence coefficient would not be $1 - \alpha$ as originally thought; that value applies only to a preplanned comparison, one planned before looking at the sample data.

One way to allow for data snooping after observing the sample data is to use a multiple-comparison procedure that has a confidence coefficient to cover all comparisons that could be done after observing the sample data. Some of these procedures are discussed in this chapter.

exploratory
hypothesis generation

The other possibility is to use data-snooping comparisons as a basis for generating **exploratory hypotheses** that must be confirmed in future experiments or studies. Here the data-snooping comparisons serve an exploratory, or hypothesis-generating, role, and inferences would not be made based on the data snoop. Further experimentation would be done to confirm (or not) the hypothesis generated in the data snoop.

Abstract of Research Study: Are Interviewers' Decisions Affected by Different Handicap Types?

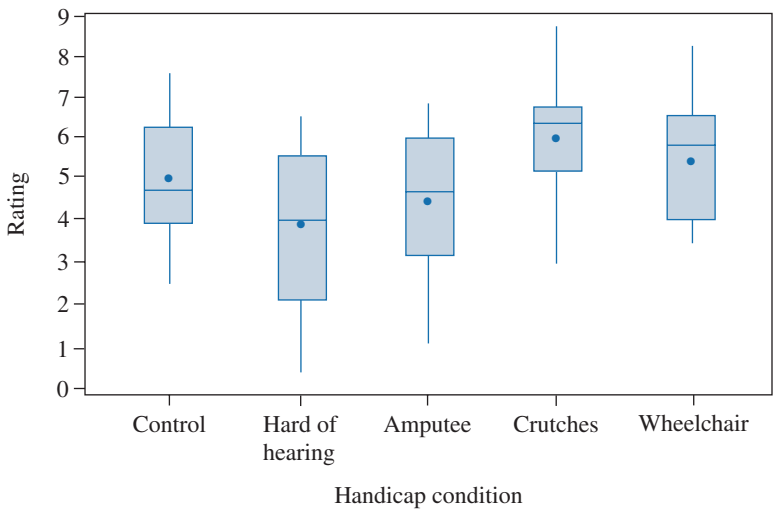
There are approximately 50 million people in the United States who report having a handicap. Furthermore, it is estimated that the unemployment rate of noninstitutionalized handicapped people between the ages of 18 and 64 is nearly double the unemployment rate of people with no impairment. Thus, it appears that people with disabilities have a more difficult time obtaining employment. One of the problems confronting people having a handicap may be a bias by employers during the employment interview.

The paper "Interviewers' Decisions Related to Applicant Handicap Type and Rater Empathy" ***(Cesare et al., 1990)***, describes a study that examines these issues. The purposes of the study were to investigate whether different types of physical handicaps produce different levels of empathy in raters and to examine if interviewers' evaluations are affected by the type of handicap of the person being interviewed.

A group of undergraduate students was randomly assigned to one of five experimental conditions that simulated an employment interview with an applicant having one of five conditions: used a wheelchair, used Canadian crutches, was hard of hearing, had a leg amputated, or was nonhandicapped (control). The researchers had a number of research questions, including the following:

1. Is there a difference in the average empathy scores of the student raters based on the type of condition viewed? (This research question could be answered using the analysis of variance procedures developed in Chapter 8.)

FIGURE 9.1
Boxplots of ratings by handicap (means are indicated by solid circles)



2. Which pairs of handicap conditions produced different average qualification scores? (The research hypothesis for analysis of variance is that there is a difference in the five treatments, but it does not address which treatments are the same or different.)
3. Is the average rating for hard-of-hearing applicants different from the average rating for applicants with mobility problems? (This research question involves comparing the average response of one treatment to the average responses of several treatments. We will define this comparison as a linear contrast in the next section.)

The researchers conducted the experiments and obtained the ratings of the applicant qualifications from 70 raters. The data are summarized in Figure 9.1. The boxplots display somewhat higher qualification scores from the raters viewing the crutches condition. The mean qualification scores for the hard of hearing and amputee conditions were somewhat smaller than those of the control and wheelchair conditions.

In the following sections, we will develop the various methodologies needed to answer the questions such as the three we have posed above. These methodologies will then be applied to the ratings data in Section 9.8.

9.2 Linear Contrasts

Before developing several different multiple-comparison procedures, we need the following notation and definitions. Consider a completely randomized design where we wish to make comparisons among the t population means $\mu_1, \mu_2, \ldots, \mu_t$. These comparisons among t population means can be written in the form

$$l = a_1\mu_1 + a_2\mu_2 + \cdots + a_t\mu_t = \sum_{i=1}^{t} a_i\mu_i$$

where the a_is are constants satisfying the property that $\sum a_i = 0$. For example, if we wanted to compare μ_1 to μ_2, we would write the linear form

$$l = \mu_1 - \mu_2$$

Note that $a_1 = 1, a_2 = -1, a_3 = a_4 = \cdots = a_t = 0$, and $\sum_i a_i = 0$. Similarly, we could compare the mean for population 1 to the average of the means for populations 2 and 3. Then l would be of the form

$$l = \mu_1 - \frac{(\mu_2 + \mu_3)}{2}$$

where $a_1 = 1, a_2 = a_3 = -\frac{1}{2}, a_4 = a_5 = \cdots = a_t = 0$, and $\sum_i a_i = 0$.

We often write the contrasts with all the a_is as integer values. We accomplish this by rewriting the a_is with a common denominator and then multipling the a_is by this common denominator. Suppose we have the following contrast in four treatment means:

$$a_1 = \frac{1}{4} \quad a_2 = \frac{-1}{6} \quad a_3 = \frac{-1}{3} \quad a_4 = \frac{1}{4}$$

The common denominator is 12, which we multiply by each of the a_is, yielding

$$a_1 = 3 \quad a_2 = -2 \quad a_3 = -4 \quad a_4 = 3$$

The two contrasts yield equivalent comparisons concerning the differences in the μs, but the integer form is somewhat easier to work with in many of our calculations.

$\hat{l}$
linear contrast

An estimate of the linear form l, designated by $\hat{l}$, is formed by replacing the μ_is in l with their corresponding sample means $\bar{y}_{i.}$. The estimate $\hat{l}$ is called a **linear contrast.**

DEFINITION 9.1

$\hat{l} = a_1\bar{y}_{1.} + a_2\bar{y}_{2.} + \cdots + a_t\bar{y}_{t.} = \sum_i a_i\bar{y}_{i.}$ is called a **linear contrast** among the t sample means and can be used to estimate $l = \sum_i a_i\mu_i$. The a_is are constants satisfying the constraint $\sum_i a_i = 0$.

The variance of the linear contrast $\hat{l}$ can be estimated as follows:

$\hat{V}(\hat{l})$

$$\hat{V}(\hat{l}) = s_W^2\left[\frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + \cdots + \frac{a_t^2}{n_t}\right] = s_W^2\sum_i \frac{a_i^2}{n_i}$$

where n_i is the number of sample observations selected from population i and s_W^2 is the mean square within samples obtained from the analysis of variance table for the completely randomized design. If all sample sizes are the same (i.e., all $n_i = n$), then

$$\hat{V}(\hat{l}) = \frac{s_W^2}{n}\sum_i a_i^2$$

Many different contrasts can be formed among the t sample means. A special set of contrasts is defined next.

DEFINITION 9.2

Two contrasts $\hat{l}_1$ and $\hat{l}_2$, where

$$\hat{l}_1 = \sum_i a_i \bar{y}_{i.} \qquad \text{and} \qquad \hat{l}_2 = \sum_i b_i \bar{y}_{i.}$$

are said to be **orthogonal** if

$$\frac{a_1 b_1}{n_1} + \frac{a_2 b_2}{n_2} + \cdots + \frac{a_t b_t}{n_t} = \sum_{i=1}^{t} \frac{a_i b_i}{n_i} = 0$$

Note: If the sample sizes are the same, then the condition becomes

$$a_1 b_1 + a_2 b_2 + \cdots + a_t b_t = \sum_{i=1}^{t} a_i b_i = 0$$

mutually orthogonal

A set of contrasts is said to be **mutually orthogonal** if all pairs of contrasts in the set are orthogonal.

EXAMPLE 9.1

Consider a completely randomized design for comparing $t = 4$ populations means, μ_1, μ_2, μ_3, and μ_4, with sample sizes $n_1 = 5$, $n_2 = 4$, $n_3 = 6$, and $n_4 = 5$. Are the following contrasts orthogonal?

$$\hat{l}_1 = \bar{y}_1 - \bar{y}_3 \quad \hat{l}_2 = \bar{y}_2 - \bar{y}_4$$

Solution We can rewrite the contrasts in the following form:

$$\hat{l}_1 = \bar{y}_1 + 0(\bar{y}_2) - \bar{y}_3 + 0(\bar{y}_4)$$

$$\hat{l}_2 = 0(\bar{y}_1) + \bar{y}_2 + 0(\bar{y}_3) - \bar{y}_4$$

Thus, we identify $a_1 = 1$, $a_2 = 0$, $a_3 = -1$, $a_4 = 0$ and $b_1 = 0$, $b_2 = 1$, $b_3 = 0$, $b_4 = -1$. It is apparent that

$$\sum_{i=1}^{4} \frac{a_i b_i}{n_i} = \frac{(1)(0)}{5} + \frac{(0)(1)}{4} + \frac{(-1)(0)}{6} + \frac{(0)(-1)}{5} = 0$$

and, hence, the contrasts are orthogonal. ■

EXAMPLE 9.2

Consider a completely randomized design for comparing $t = 4$ populations means, μ_1, μ_2, μ_3, and μ_4, with sample sizes $n_1 = 5$, $n_2 = 4$, $n_3 = 6$, and $n_4 = 5$. Are the following contrasts orthogonal?

$$\hat{l}_1 = \bar{y}_1 - \bar{y}_3 \qquad \hat{l}_2 = \bar{y}_1 + \bar{y}_2 + \bar{y}_3 - 3(\bar{y}_4)$$

Solution We can rewrite the contrasts in the following form:

$$\hat{l}_1 = \bar{y}_1 + 0(\bar{y}_2) - \bar{y}_3 + 0(\bar{y}_4)$$

$$\hat{l}_2 = \bar{y}_1 + \bar{y}_2 + \bar{y}_3 - 3(\bar{y}_4)$$

Thus, we identify $a_1 = 1, a_2 = 0, a_3 = -1, a_4 = 0$ and $b_1 = 1, b_2 = 1, b_3 = 1, b_4 = -3$. The evaluation of orthoganality is as follows

$$\sum_{i=1}^{4} \frac{a_i b_i}{n_i} = \frac{(1)(1)}{5} + \frac{(0)(1)}{4} + \frac{(-1)(1)}{6} + \frac{(0)(-3)}{5} = \frac{1}{5} - \frac{1}{6} = \frac{1}{30}.$$

Thus, the contrasts are not orthogonal. Note that if the sample sizes were all equal—say, $n_i = 5$ for all i—then

$$\sum_{i=1}^{4} \frac{a_i b_i}{n_i} = \frac{(1)(1)}{5} + \frac{(0)(1)}{5} + \frac{(-1)(1)}{5} + \frac{(0)(-3)}{5} = \frac{1}{5} - \frac{1}{5} = 0$$

and the two contrasts would have been orthogonal. ■

$t - 1$ contrasts

The concept of orthogonality between linear contrasts is important because if two contrasts are orthogonal, then one contrast conveys no information about the other contrast. We will demonstrate that $t - 1$ orthogonal contrasts can be formed using the t sample means, $\bar{y}_{i.}$s. These ***t* − 1 contrasts*** form a set of mutually orthogonal contrasts. (An easy way to remember $t - 1$ is to refer to the number of degrees of freedom associated with the treatment (between-sample) source of variability in the AOV table.) In addition, it can be shown that the sums of squares for the $t - 1$ contrasts will add up to the treatment (between-sample) sum of squares. Mutual orthogonality is desirable because it leads to the independence of the $t - 1$ sums of squares associated with the $t - 1$ orthogonal contrasts. Thus, we can take the $t - 1$ degrees of freedom associated with the treatment sum of squares that describe any differences among the treatment means and break them into $t - 1$ independent explanations of how the treatment means may differ. We will now further develop these ideas and illustrate the concepts with an example.

A sum of squares associated with a treatment contrast is calculated to indicate the amount of variation in the treatment means that can be *explained* by that particular contrast. For each contrast $\hat{l} = \sum_{i=1}^{t} a_i \bar{y}_{i.}$, we can calculate a sum of squares associated with that contrast (SSC):

$$\text{SSC} = \frac{(\sum_{i=1}^{v} a_i \bar{y}_i)^2}{\sum_{i=1}^{v} (a_i^2/n_i)} = \frac{(\hat{l})^2}{\sum_{i=1}^{v} (a_i^2/n_i)}$$

When the sample sizes are equal, $n_1 = n_2 = \cdots = n_t = n$, this formula simplifies to

$$\text{SSC} = \frac{n(\hat{l})^2}{\sum_{i=1}^{t} a_i^2}$$

Associated with each such sum of squares is 1 degree of freedom. Thus, we can obtain $t - 1$ orthogonal contrasts such that the sum of squares treatment, which has $t - 1$ degrees of freedom, equals the total of the $t - 1$ sum of squares associated with each of the contrasts. The following example illustrates these calculations.

EXAMPLE 9.3

Various agents are used to control weeds in crops. Of particular concern is the overusage of chemical agents. Although effective in controlling weeds, these agents may also drain into the underground water system and cause health problems. Thus, several new biological weed agents have been proposed to eliminate the contamination problem present in chemical agents. Researchers conducted a study of biological agents to assess their effectiveness in comparison to the chemical weed agents. The study consisted of a control (no agent), two biological agents (Bio1 and Bio2), and two chemical agents (Chm1 and Chm2). Thirty 1-acre plots of land were planted with hay. Six plots were randomly assigned to receive one of the five treatments. The hay was harvested, and the total yield in tons per acre was recorded for each plot. The data are given in Table 9.1.

TABLE 9.1
Summary statistics for Example 9.3

Agent	1	2	3	4	5
Type	None	Bio1	Bio2	Chm1	Chm2
$\bar{y}_{i.}$	1.175	1.293	1.328	1.415	1.500
s_i	.1204	.1269	.1196	.1249	.1265
n_i	6	6	6	6	6

Determine four orthogonal contrasts, and demonstrate that the total of the four sums of squares associated with the four contrasts equals the between-samples (treatment) sum of squares.

Solution An analysis of variance was conducted on these data yielding the results summarized in the AOV table given in Table 9.2.

TABLE 9.2
AOV table for Example 9.3

Source	df	SS	MS	F	p-value
Treatment	4	.3648	.0912	5.96	.0016
Error	25	.3825	.0153		
Total	29	.7473			

From the AOV table, we have that $SS_{Trt} = .3648$. We will now construct four orthogonal contrasts in the five treatment means and demonstrate that SS_{Trt} can be partitioned into four terms, each representing a 1 degree of freedom sum of squares associated with a particular contrast. Table 9.3 contains the coefficient and sum of squares for each of the four contrasts.

TABLE 9.3
Sum of squares computations for weed control experiment

	Treatment							
Contrast	1(Cntrl) a_1	2(Bio1) a_2	3(Bio2) a_3	4(Chm1) a_4	5(Chm2) a_5	$\sum_{i=1}^{5} a_i^2$	$\hat{l}$	SSC_i
Control vs. Agents	4	−1	−1	−1	−1	20	−.836	.2097
Biological vs. Chemical	0	1	1	−1	−1	4	−.294	.1297
Bio1 vs. Bio2	0	1	−1	0	0	2	−.035	.0037
Chm1 vs. Chm2	0	0	0	1	−1	2	−.085	.0217
$\bar{y}_{i.}$	1.175	1.293	1.328	1.415	1.500			.3648

To illustrate the calculations involved in Table 9.3, we will compute the sum of squares associated with the first contrast, control versus agents. First, note that the contrast represents a comparison of the yield for the control treatment versus the average yield of the four active agents. We initially would have written this contrast as

$$l = \mu_1 - \frac{(\mu_2 + \mu_3 + \mu_4 + \mu_5)}{4}$$
$$= (1)\,\mu_1 + \left(\frac{-1}{4}\right)\mu_2 + \left(\frac{-1}{4}\right)\mu_3 + \left(\frac{-1}{4}\right)\mu_4 + \left(\frac{-1}{4}\right)\mu_5$$

However, we can multiply each coefficient by 4 and change the coefficients from

$$a_1 = 1 \quad a_2 = \frac{-1}{4} \quad a_3 = \frac{-1}{4} \quad a_4 = \frac{-1}{4} \quad a_5 = \frac{-1}{4}$$

to

$$a_1 = 4 \quad a_2 = -1 \quad a_3 = -1 \quad a_4 = -1 \quad a_5 = -1$$

Next, we calculate

$$\sum_{i=1}^{5} a_i^2 = (4)^2 + (-1)^2 + (-1)^2 + (-1)^2 + (-1)^2 = 20$$

and

$$\hat{l} = (4)(1.175) + (-1)(1.293) + (-1)(1.328) + (-1)(1.415) + (-1)(1.500) = -.836$$

Finally, we can obtain the sum of squares associated with the contrast from

$$\text{SSC}_1 = \frac{(\hat{l})^2}{\sum_{i=1}^{5}(a_i^2/n_i)} = \frac{n(\hat{l})^2}{\sum_{i=1}^{5} a_i^2} = \frac{6(-.836)^2}{20} = .2097$$

The remaining three sums of squares are calculated in a similar fashion. From Table 9.3, we thus obtain

$$\text{SSC}_1 + \text{SSC}_2 + \text{SSC}_3 + \text{SSC}_4 = .2097 + .1297 + .0037 + .0217 = .3648 = \text{SS}_{\text{Trt}}$$ ■

EXAMPLE 9.4

Refer to Example 9.3. Verify that the four contrasts in Table 9.3 are mutually orthogonal.

Solution Identify the four contrasts in Table 9.3 by $\hat{l}_1$ is Control vs. Agents, $\hat{l}_2$ is Biological vs. Chemical, $\hat{l}_3$ is Bio1 vs. Bio2, and $\hat{l}_4$ is Chm1 vs. Chm2. Note that the sample sizes are equal, so we need to verify that $\sum_{i=1}^{5} a_i b_i = 0$ for the six pairs of contrasts. (See Table 9.4.)

TABLE 9.4 Verification of orthogonality

Contrast	Verification of Orthogonality
$\hat{l}_1$ and $\hat{l}_2$	$\sum_{i=1}^{5} a_i b_i = (4)(0) + (-1)(1) + (-1)(1) + (-1)(-1) + (-1)(-1) = 0$
$\hat{l}_1$ and $\hat{l}_3$	$\sum_{i=1}^{5} a_i b_i = (4)(0) + (-1)(1) + (-1)(-1) + (-1)(0) + (-1)(0) = 0$
$\hat{l}_1$ and $\hat{l}_4$	$\sum_{i=1}^{5} a_i b_i = (4)(0) + (-1)(0) + (-1)(0) + (-1)(1) + (-1)(-1) = 0$
$\hat{l}_2$ and $\hat{l}_3$	$\sum_{i=1}^{5} a_i b_i = (0)(0) + (1)(1) + (1)(-1) + (-1)(0) + (-1)(0) = 0$
$\hat{l}_2$ and $\hat{l}_4$	$\sum_{i=1}^{5} a_i b_i = (0)(0) + (1)(0) + (1)(0) + (-1)(1) + (-1)(-1) = 0$
$\hat{l}_3$ and $\hat{l}_4$	$\sum_{i=1}^{5} a_i b_i = (0)(0) + (1)(0) + (-1)(0) + (0)(1) + (0)(-1) = 0$

■

Example 9.3 illustrated how we can decompose differences in the treatment means into individual contrasts that represent various comparisons of the treatment means. After defining the contrasts and obtaining their estimates and sums of squares, we need to determine which of the contrasts are significantly different from zero. A value of zero for a contrast would indicate that the difference in the means represented by the contrast does not exist. For example, if our contrast l_1 (control versus agents) was determined to be zero, then we would conclude that the average yield on plots assigned no agent (control) was equal to the average yield across all plots having one of the four agents. We will now present a test of the hypothesis that a contrast $l = \sum_{i=1}^{t} a_i \mu_i$ is different from zero. Our test will be a variation of the F test from AOV. Because the sum of squares associated with a

contrast has 1 degree of freedom, its mean square is the same as its sum of squares. The test statistic is simply

$$F = \frac{\text{SSC}}{\text{MS}_{\text{Error}}} = \frac{\text{SSC}}{s_W^2}$$

The test procedure is summarized here.

F Test for Contrasts

H_0: $l = a_1\mu_1 + a_2\mu_2 + \cdots + a_t\mu_t = 0$

H_a: $l = a_1\mu_1 + a_2\mu_2 + \cdots + a_t\mu_t \neq 0$

T.S.: $F = \dfrac{\text{SSC}}{\text{MS}_{\text{Error}}}$

R.R.: For a specified value of α, reject H_0 if F exceeds the tabled F value (Table 8 in the Appendix) for the specified α, $\text{df}_1 = 1$, and $\text{df}_2 = n_T - t$.

Check assumptions and draw conclusions.

EXAMPLE 9.5

Refer to Example 9.3. The researchers were very interested in determining whether the biological agents would perform as well as the chemical agents. Is there a significant difference between the control treatment and the four active agents for weed control with respect to their effect on average hay production? Test each of the four contrasts for significance.

Solution From the table of summary statistics in Example 9.3, the sample standard deviations are nearly equal. Thus, we have very little reason to suspect that the five population variances are unequal. The AOV table in Example 9.3 has a p-value of .0016. Thus, we have a very strong rejection of H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. We thus conclude that there are significant (p-value = .0016) differences in the five treatment means. We can investigate the types of differences in these means using the four contrasts that we constructed in Example 9.3. The four test statistics are computed here with $F_i = \text{SSC}_i/\text{MS}_{\text{Error}}$:

$$F_1 = \frac{.2097}{.0153} = 13.71 \quad F_2 = \frac{.1297}{.0153} = 8.48 \quad F_3 = \frac{.0037}{.0153} = 0.24$$

$$F_4 = \frac{.0217}{.0153} = 1.42$$

From Table 8 in the Appendix, with $\alpha = .05$, $\text{df}_1 = 1$, and $\text{df}_2 = 30 - 5 = 25$, we obtain $F_{.05,1,25} = 4.24$. Thus, we conclude that contrasts l_1 and l_2 were significantly different from zero but that contrasts l_3 and l_4 were not significantly different from zero. Using contrast l_1, we could thus conclude that the mean yields from plots using a weed control agent were significantly higher than the mean yields from plots on which no agent was used. From contrast l_2, we infer that the mean yields from fields using biological agents for weed control would tend to be lower than the mean yields from those using chemical agents. However, we would need to investigate the size of the differences in the mean yields to determine whether the differences were of economical importance rather than just statistically significantly different. If the differences were economically significant, the ecological gains from using the biological agents might justify their use in place of chemical agents. ■

When we select contrasts for a study, the goal is not to obtain a set of orthogonal contrasts that yield a decomposition of the sum of squares treatment into $t - 1$ components. Rather, the goal is to obtain contrasts of the treatment means that will elicit a clear explanation of the pattern of differences in the treatment means of most benefit to the researcher. The mutual orthogonality of the contrasts is somewhat of a fringe benefit of the selection process. For example, in the analysis of the weed agents, we may have also been interested in comparing the control treatment to the average of the two biological agents. This contrast would not have been orthogonal to several of the contrasts we had already designed. We could have still used this contrast and tested its significance using the experimental data. The choice of which contrasts to evaluate should be determined by the overall goals of the experimenter and not by orthogonality.

One problem we do encounter when testing a number of contrasts is referred to as multiple comparisons. When we have tested several contrasts, each with a Type I error rate of α, the chance of at least one Type I error occurring during the several tests becomes somewhat larger than α. In the next section, we will address this difficulty.

9.3 Which Error Rate Is Controlled?

individual comparisons Type I error rate

experimentwise Type I error rate

An experimenter wishes to compare t population (treatment) means using m contrasts. Each of the m contrasts can be tested using the F test we introduced in the previous section. Suppose each of the contrasts is tested with the same value of α, which we will denote as α_I, called the **individual comparisons Type I error rate**. Thus, we have an α_I chance of making a Type I error on each of the m tests. We need to also consider the probability of falsely rejecting at least one of the m null hypotheses, called the **experimentwise Type I error rate** and denoted by α_E. The value of α_E takes into account that we are conducting m tests, each having an α_I chance of making a Type I error. Now, if MS_{Error} has an infinite number of degrees of freedom (so the tests are independent), then when all m null hypotheses are true, the probability of falsely rejecting at least one of the m null hypotheses can be shown to be $\alpha_E = 1 - (1 - \alpha_I)^m$. Table 9.5 contains values of α_E for various values of m and α_I. We can observe from Table 9.5 that as the number of tests m increases for a given value of α_I, the probability of falsely rejecting H_0 on at least one of the m tests,

TABLE 9.5
A comparison of the experimentwise error rate, α_E, for m independent contrasts among $t(t > m)$ sample means

m, Number of Contrasts	α_I, Probability of a Type I Error on an Individual Test		
	.10	.05	.01
1	.100	.050	.010
2	.190	.097	.020
3	.271	.143	.030
4	.344	.185	.039
5	.410	.226	.049
.	.	.	.
.	.	.	.
.	.	.	.
10	.651	.401	.096

α_E, becomes quite large. For example, if an experimenter wanted to compare $t = 20$ population means by using $m = 10$ orthogonal contrasts, the probability of falsely rejecting H_0 on at least one of the t tests could be as high as .401 when each individual test was performed with $\alpha_I = .05$.

In any practical problem, the degrees of freedom for MS_{Error} will not be infinite, and, hence, the tests will not be independent. Thus, the relationship between α_E and α_I is not generally as described in Table 9.5. It is difficult to obtain an expression equivalent to $\alpha_E = 1 - (1 - \alpha_I)^m$ for comparisons made with tests that are not independent. However, it can be shown that for most of the types of comparisons we will be making among the population means, the following upper bound exists for the experimentwise error rate:

$$\alpha_E \leq 1 - (1 - \alpha_I)^m$$

Thus, we know the largest possible value for α_E when we set the value of α_I for each of the individual tests. Suppose, for example, that we wish the experimentwise error rate for $m = 8$ contrasts among $t = 20$ population means to be at most .05. What value of α_I must we use on the m tests to achieve an overall error rate of $\alpha_E = .05$? We can use the previous upper bound to determine that if we select

$$\alpha_I = 1 - (1 - \alpha_E)^{1/m} = 1 - (1 - .05)^{1/8} = .0064$$

then we will have $\alpha_E \leq .05$. The only problem is that this procedure may be very conservative with respect to the experimentwise error rate, and, hence, an inflated probability of Type II error may result.

We will now consider a method that will work for any set of m tests and is much easier to apply in obtaining an upper bound on α_E. The results of Table 9.5 are disturbing when we are conducting a number of tests. The chance of making at least one Type I error may be considerably larger than the selected individual error rates. This could lead us to question significant results when they appear in our analysis of experimental results. The problem can be alleviated somewhat by *controlling the experimentwise error rate* α_E rather than the *individual error rate* α_I. We need to select a value of α_I that will provide us with an acceptable value for α_E.

Bonferroni inequality

The **Bonferroni inequality** provides us with a method for selecting α_I so that α_E is bounded below a specified value. This inequality states that the overall Type I error rate α_E is less than or equal to the sum of the individual error rates for the m tests. Thus, when each of the m tests has the same individual error rate, α_I, the Bonferroni inequality yields

$$\alpha_E \leq m\alpha_I$$

If we wanted to guarantee that the chance of a Type I error was at most α, we could select

$$\alpha_I = \frac{\alpha}{m}$$

for each of the m tests. Then

$$\alpha_E \leq m\alpha_I = m\left(\frac{\alpha}{m}\right) = \alpha$$

The experimentwise error rate is thus less than or equal to our specified value. Just as we mentioned earlier, this procedure may be very conservative with respect to the experimentwise error rate, and, hence, an inflated probability of Type II error may result.

EXAMPLE 9.6

Refer to Example 9.5, where we constructed $m = 4$ contrasts (comparisons) among the $t = 5$ treatment means. If we wanted to control the experimentwise error rate at a level of $\alpha_E = .05$, then we would take

$$\alpha_I = \frac{0.5}{4} = .0125$$

The critical value for the F tests would be $F_{.0125,\,1,\,25} = 7.24$, which can be obtained using the R function ***qf***(1 −.0125, 1, 25). We would then reject H_0 if $F_i = \text{SSC}_i/\text{MS}_{\text{Error}} \geq 7.24$. The Bonferroni critical value, 7.24, is much larger than $F_{.05,\,1,\,25} = 4.24$, the critical value obtained ignoring the impact of multiple testing. Thus, the Bonferroni procedure will potentially lead to fewer contrasts being declared significantly different from 0. From Example 9.5, the four F ratios were

$$F_1 = 13.71 \qquad F_2 = 8.48 \qquad F_3 = 0.24 \qquad F_4 = 1.42$$

Using the Bonferroni procedure, we would declare contrast l_1 and l_2 significantly different from 0 because their F ratios are greater than 7.24. Using the Bonferroni test procedure, we are assured that the chance of making at least one Type I error during the four tests is at most .05. Using $\alpha = .05$ for each of the four procedures would not have allowed us to assess the exact probability of making a Type I error among the four comparisons. However, this value would have been considerably larger than .05, possibly as large as .20. ■

The Bonferroni procedure gives us a method for evaluating a small number of contrasts that were selected prior to observing the data, while preserving a selected experimentwise Type I error rate. In some experimental settings, the researcher may want to test a large number of contrasts. A procedure proposed by Scheffé (1953) can be used to make all possible comparisons among t population means. Scheffé's procedure provides the selected experimentwise error rate for any number of contrasts, whereas the Bonferroni procedure only sets an upper bound on the experimentwise error rate.

9.4 Scheffé's *S* Method

The Scheffé procedure is a very general procedure that can be used to test the significance of all possible contrasts among t population means, while maintaining the selected experimentwise error rate. Because the procedure can be applied to an unlimited number of contrasts, it is a very conservative procedure (less sensitive) than many other procedures for testing contrasts. The other procedures, which will be introduced later in this chapter, are developed for specific comparisons such as comparing all pairs of means or comparing $t - 1$ treatments to a control. Thus, these procedures have a much smaller number of tests for which the experimentwise error rate needs to be controlled than the unlimited number being controlled by the Scheffé procedure.

Scheffé's *S* Method for Multiple Comparisons

1. Consider any linear comparison among the t population means of the form
$$l = a_1\mu_1 + a_2\mu_2 + \cdots + a_t\mu_t$$
We wish to test the null hypothesis
$$H_0: l = 0$$
against the alternative
$$H_a: l \neq 0$$
2. The test statistic is
$$\hat{l} = a_1\bar{y}_{1.} + a_2\bar{y}_{2.} + \cdots + a_t\bar{y}_{t.}$$
3. Let
$$S = \sqrt{\hat{V}(\hat{l})}\sqrt{(t-1)F_{\alpha,\mathrm{df}_1,\mathrm{df}_2}}$$
where, from Section 9.2,
$$\hat{V}(\hat{l}) = s_W^2\sum_i \frac{a_i^2}{n_i}$$
t is the total number of population means and $F_{\alpha,\mathrm{df}_1,\mathrm{df}_2}$ is the upper-tail critical value of the F distribution for the specified value of α, with $\mathrm{df}_1 = t - 1$ and df_2 the degrees of freedom for s_W^2.
4. For a specified value of α, we reject H_0 if $|\hat{l}| > S$.
5. The error rate that is controlled is an *experimentwise error rate*. If we consider all imaginable contrasts, the probability of observing an experiment with one or more contrasts falsely declared to be significant is designated by α.

EXAMPLE 9.7

Refer to Example 9.5. We defined four contrasts in the $t = 5$ treatment means in an attempt to investigate the differences in the average hay production on fields treated with either the control or one of the four weed agents. Use the sample data and Scheffé's procedure to determine which if any of the four contrasts are significantly different from zero. Use $\alpha = .05$.

Solution The four contrasts of interest are given in Table 9.6 along with their estimates. To illustrate the calculations involved in Table 9.6, we will compute the value of S for the first contrast, control vs. agents. To compute
$$S = \sqrt{\hat{V}(\hat{l})}\sqrt{(t-1)F_{\alpha,\mathrm{df}_1,\mathrm{df}_2}}$$

TABLE 9.6 Computations for Scheffé procedure in weed control experiment

	Treatment									
Contrast	**Control** a_1	**Bio1** a_2	**Bio2** a_3	**Chm1** a_4	**Chm2** a_5	$\sum a_i^2/n_i$	$\hat{l}$	$\hat{V}(\hat{l})$	S	**Conclusion**
Control vs. agents	4	−1	−1	−1	−1	20/6	−.836	.0510	.750	Significant
Biological vs. chemical	0	1	1	−1	−1	4/6	−.294	.0102	.336	Not significant
Bio1 vs. Bio2	0	1	−1	0	0	2/6	−.035	.0051	.237	Not significant
Chm1 vs. Chm2	0	0	0	1	−1	2/6	−.085	.0051	.237	Not significant

we must first calculate $\hat{V}(\hat{l})$. Using the formula

$$\hat{V}(\hat{l}) = s_W^2 \sum_{i=1}^{t} \frac{a_i^2}{n_i}$$

with all samples sizes equal to 6 and $s_W^2 = .0153$, we have

$$\hat{V}(\hat{l}) = .0153\left(\frac{(4)^2}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) = .0153\,\frac{20}{6} = .0510$$

From Table 8 in the Appendix for $\alpha = .05$, $df_1 = t - 1 = 4$, and $df_2 = 25$ (the degrees of freedom for s_W^2), $F_{.05,4,25} = 2.76$. The computed value of S is then

$$S = \sqrt{.0510}\sqrt{4(2.76)} = (.2258)(3.323) = .750$$

Because the absolute value of $\hat{l}$ is $|-.836| = .836$, which exceeds .750, we have significant evidence ($\alpha = .05$) to indicate that the average hay production from the fields treated with a weed agent exceeds the average yield in the fields having no treatment for weeds. The calculations for the other three contrasts are summarized in Table 9.6. Note that the value of S changes for the different contrasts. In our example, the only contrast significantly different from zero was the first contrast. The remaining three contrasts were not significant at the $\alpha = .05$ level. These conclusions are different from the conclusions we reached in Example 9.5, where we found that the second contrast was also significantly different from zero. The reason for the difference in the conclusions is that the Scheffé procedure controls the experimentwise Type I error rate at level .05, whereas in Example 9.5 we only control the individual comparison rate at level .05. ■

Scheffé's confidence interval

Scheffé's method can also be used for constructing a simultaneous confidence interval for all possible (not necessarily pairwise) contrasts using the t treatment means. In particular, there is a probability equal to $1 - \alpha$ that all possible comparisons of the form $l = \sum a_i \mu_i$, where $\sum a_i = 0$, will be encompassed by intervals of the form

$$(\hat{l} - S, \hat{l} + S)$$

9.5 Tukey's *W* Procedure

Studentized range distribution

Tukey (1953) proposed a procedure that makes use of the **Studentized range distribution.** When more than two sample means are being compared, to test the largest and smallest sample means, we could use the test statistic

$$\frac{\bar{y}_{\text{largest}} - \bar{y}_{\text{smallest}}}{s_p\sqrt{1/n}}$$

where n is the number of observations in each sample and s_p is a pooled estimate of the common population standard deviation σ. This test statistic is very similar to that for comparing two means, but it does not possess a t distribution. One reason it does not is that we have waited to determine which two sample means (and hence population means) we would compare until we observed the largest and smallest sample means. This procedure is quite different from that of specifying H_0: $\mu_1 - \mu_2 = 0$, observing $\bar{y}_{1.}$ and $\bar{y}_{2.}$, and forming a t statistic.

The quantity

$$\frac{\bar{y}_{\text{largest}} - \bar{y}_{\text{smallest}}}{s_p\sqrt{1/n}}$$

follows a Studentized range distribution. We will not discuss the properties of this distribution but will illustrate its use in Tukey's multiple-comparison procedure.

Tukey's *W* Procedure

1. Rank the t sample means.
2. Two population means μ_i and μ_j are declared different if

   $$|\bar{y}_{i.} - \bar{y}_{j.}| \geq \boldsymbol{W}$$

   where

   $$W = \boldsymbol{q_\alpha(t, \nu)}\sqrt{\frac{s_W^2}{n}}$$

   s_W^2 is the mean square within samples based on ν degrees of freedom, $q_\alpha(t, \nu)$ is the **upper-tail critical value of the Studentized range** for comparing t different populations, and n is the number of observations in each sample. A discussion follows showing how to obtain values of $q_\alpha(t, \nu)$ by referring to Table 10 in the Appendix or using the R function ***q*tukey**$(1 - \alpha, t, \nu)$.
3. The error rate that is controlled is an **experimentwise error rate.** Thus, the probability of observing an experiment with one or more pairwise comparisons falsely declared to be significant is specified at α.

W

***q*$_\alpha(t, \nu)$**

upper-tail critical value of the Studentized range

experimentwise error rate

We can obtain values of $\boldsymbol{q_\alpha}(t, \nu)$ from Table 10 in the Appendix. Values of ν are listed along the left column of the table with values of t across the top row. Upper-tail values for the Studentized range are then presented for $\alpha = .05$ and .01. For example, in comparing 10 population means based on 9 degrees of freedom for s_W^2, the .05 upper-tail critical value of the Studentized range is $\boldsymbol{q}_{.05}(10, 9) = 5.74$.

EXAMPLE 9.8

Refer to the data of Example 9.3. Use Tukey's W procedure with $\alpha = .05$ to make pairwise comparisons among the five population means.

Solution Step 1 is to rank the sample means from smallest to largest, to produce the following table:

Agent	1	2	3	4	5
$\bar{y}_{i.}$	1.175	1.293	1.328	1.415	1.500

For the experiment described in Example 9.3, we have

$t = 5$ (we are making pairwise comparisons among five means)
$\nu = 25$ (s_W^2 had degrees of freedom equal to df_{Error} in the AOV)
$\alpha = .05$ (we specified α_E, the experimentwise error rate at .05)
$n = 6$ (there were six plots randomly assigned to each of the agents)

We find in Table 10 of the Appendix that

$$q_\alpha(t, \nu) = q_{.05}(5, 25) \approx 4.158$$

Alternatively,

$$\boldsymbol{q}\textbf{tukey}(.95, 5, 25) = 4.153$$

The absolute value of each difference in the sample means $|\bar{y}_i - \bar{y}_i|$ must then be compared to

$$W = q_\alpha(t, \nu)\sqrt{\frac{s_W^2}{n}} = 4.153\sqrt{\frac{.0153}{6}} = .2097$$

Next, compute the difference in the sample means

$$\bar{y}_{\text{largest}} - \bar{y}_{\text{smallest}}$$

If this difference is greater than W, we declare the corresponding population means significantly different from each other. Next, we compute the difference in the sample means

$$\bar{y}_{\text{2nd largest}} - \bar{y}_{\text{smallest}}$$

and compare the difference to W. We continue to compute differences with $\bar{y}_{\text{smallest}}$

$$\bar{y}_{\text{3rd largest}} - \bar{y}_{\text{smallest}}$$

and so on until we find either that all differences in the sample means involving $\bar{y}_{\text{smallest}}$ exceed W (and hence the corresponding population means are different) or that a difference in the sample means involving $\bar{y}_{\text{smallest}}$ is less that W. In the latter case, we stop and make no further comparisons with $\bar{y}_{\text{smallest}}$. For our data, comparisons with $\bar{y}_{\text{smallest}}$, $\bar{y}_1$ yield the following results:

Comparison	Conclusion
$\bar{y}_{\text{largest}} - \bar{y}_{\text{smallest}} = \bar{y}_{5.} - \bar{y}_{1.} = .325$	$>W$; proceed
$\bar{y}_{\text{2nd largest}} - \bar{y}_{\text{smallest}} = \bar{y}_{4.} - \bar{y}_{1.} = .240$	$>W$; proceed
$\bar{y}_{\text{3rd largest}} - \bar{y}_{\text{smallest}} = \bar{y}_{3.} - \bar{y}_{1.} = .153$	$<W$; stop

To summarize our results we make the following diagram:

Agent 1 2 3 4 5

Next, comparison with $\bar{y}_{\text{2nd smallest}}$, which is $\bar{y}_{2.}$, yields

Comparison	Conclusion
$\bar{y}_{5.} - \bar{y}_{2.} = .207$	$<W$; stop

Agent 1 2 3 4 5

Similarly, comparisons of $\bar{y}_{5.}$ with $\bar{y}_{3.}$ and $\bar{y}_{4.}$ yield

Agent 1 2 3 4 5

Combining our results, we obtain

Agent 1 2 3 4 5

which simplifies to

Agent 1 2 3 4 5

All populations not underlined by a common line have population means that are significantly different from each other; that is, μ_4 and μ_5 are significantly different from μ_1. No other pairs of means are significantly different. ■

A limitation of Tukey's procedure is the requirement that all the sample means be based on the same number of data values. Tukey (1953) and Kramer (1956) independently proposed an approximate procedure in the case of unequal sample sizes. In place of Tukey's W, use

$$W^* = \frac{q_\alpha(t, \nu)}{\sqrt{2}}\sqrt{s_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

to compare population means μ_i and μ_j, where n_i and n_j are the corresponding sample sizes. This procedure, Tukey–Kramer, is approximate because $\alpha_E \leq \alpha$, whereas when $n_1 = n_2 = \cdots = n_t$, $\alpha_E = \alpha$.

simultaneous confidence interval

Tukey's procedure can also be used to construct confidence intervals for comparing two means. Tukey's procedure enables us to construct **simultaneous confidence intervals** for all pairs of treatment differences. For a specified α level from which we compute W, the overall probability is $1 - \alpha$ that all differences $\mu_i - \mu_j$ will be included in an interval of the form

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm W$$

that is, the probability is $1 - \alpha$ that all the intervals $(\bar{y}_{i.} - \bar{y}_{j.}) \pm W$ include the corresponding population differences $\mu_i - \mu_j$.

EXAMPLE 9. 9

Refer to Example 9.8. Construct 95% Tukey confidence intervals on the difference in all treatment means.

Solution From Example 9.8, we have $W = q_{.05}\sqrt{\frac{s_W^2}{n}} = 4.153\sqrt{\frac{.0153}{6}} = .2097$. Thus, the confidence intervals for $\mu_i - \mu_j$ will have the form $(\bar{y}_{i.} - \bar{y}_{j.}) \pm .2097$. For example, the 95% confidence interval for $\mu_3 - \mu_1$ would be $1.328 - 1.175 \pm .2097$—that is, $(-.057, .363)$. The remaining confidence intervals are given in Table 9.7.

TABLE 9.7
Confidence intervals for Example 9.9

Difference in Means	95% C.I. for Difference	Difference in Means	95% C.I. for Difference
$\mu_2 - \mu_1$	$(-.092, .328)$	$\mu_4 - \mu_2$	$(-.088, .332)$
$\mu_3 - \mu_1$	$(-.057, .363)$	$\mu_5 - \mu_2$	$(-.003, .417)$
$\mu_4 - \mu_1$	$(.030, .450)$	$\mu_4 - \mu_3$	$(-.123, .297)$
$\mu_5 - \mu_1$	$(.115, .535)$	$\mu_5 - \mu_3$	$(-.038, .382)$
$\mu_3 - \mu_2$	$(-.175, .245)$	$\mu_5 - \mu_4$	$(-.125, .295)$

From Table 9.7, we can conclude that the only pairs of population means that are significantly different are (μ_4, μ_1) and (μ_5, μ_1). The confidence intervals for both pairs do not contain 0 whereas 0 is contained in the remaining eight confidence intervals. Recall that if 0 is contained in the confidence interval, then we cannot reject the null hypothesis H_0: $\mu_i - \mu_j = 0$, the hypothesis that the treatment means μ_i and μ_j are equal. ■

9.6 Dunnett's Procedure: Comparison of Treatments to a Control

In many studies and experiments, the researchers will include a control treatment for comparison purposes. There are many types of controls, but generally the control serves as a standard to which the other treatments may be compared. For example, in many situations, the conditions under which the experiment is run may have such a strong effect on the response variable that generally effective treatments will not produce a favorable response in the experiment. For example, if the insect population is too dense, most insecticides used at a reasonable level would not provide a noticeable reduction in the insect population. Thus, a control spray with no active ingredient would reveal the level of insects in the sprayed region. A second situation in which a control is useful is when the experimental participants generate a favorable response whenever any reasonable treatment is applied; this is referred to as the **placebo effect.** In this type of study or experiment, the participants randomly assigned to the control treatment are handled in exactly the same manner as the participants receiving active treatments. In most clinical trials and experiments used to evaluate new drugs or medical treatments, a placebo treatment is included so as to determine the size of the placebo effect. Finally, a control may represent the current method or standard procedure to which any new procedures would be compared.

placebo effect

In experiments in which a control is included, the researchers want to determine whether the mean responses for the active treatments differ from the mean response for the control. Dunnett (1955) developed a procedure for comparisons to a control that controls the experimentwise Type I error rate. This procedure compares each treatment mean to the mean for the control, $\bar{y}_c$, by comparing the difference in the sample means, $\bar{y}_i - \bar{y}_c$, to the critical difference

$$D = d_\alpha(k, \nu)\sqrt{\frac{2s_W^2}{n}}$$

where $n_c = n_1 = \cdots n_{t-1} = n$. The Dunnett procedure requires equal sample sizes, $n_i = n_c$. The values for $d_\alpha(k, \nu)$ are given in Table 11 in the Appendix. Dunnett (1964) describes adjustments to the values in Table 11 for the case of unequal n_i. The comparison can be either one-sided or two-sided, as is summarized here.

Dunnett's Procedure

1. For a specified value of α_E, Dunnett's D value for comparing μ_i to μ_c, the control mean, is

   $$D = d_\alpha(k, \nu)\sqrt{\frac{2s_W^2}{n}}$$

   where n is the common sample size for the treatment groups (including the control); $k = t - 1$ is the number of noncontrol treatments; α is the desired experimentwise error rate; s_W^2 is the mean square within samples; ν is the degrees of freedom associated with s_W^2; and $d_\alpha(k, \nu)$ is the critical Dunnett value (Table 11 of the Appendix).
2. For the two-sided alternative H_a: $\mu_i \neq \mu_c$, we declare μ_i different from μ_c if

   $$|\bar{y}_i - \bar{y}_c| \geq D$$

where the value of $d_\alpha(k, v)$ is the two-sided value in Table 11 in the Appendix.

3. For the one-sided alternative H_a: $\mu_i > \mu_c$, we declare μ_i greater than μ_c if

$$(\bar{y}_i - \bar{y}_c) \geq D$$

where the value of $d_\alpha(k, v)$ is the one-sided value in Table 11 in the Appendix.

4. For the one-sided alternative H_a: $\mu_i < \mu_c$, we declare μ_i less than μ_c if

$$(\bar{y}_i - \bar{y}_c) \leq -D$$

where the value of $d_\alpha(k, v)$ is the one-sided value in Table 11 in the Appendix.

5. The Type I error rate that is controlled is an *experimentwise error rate*. Thus, the probability of observing an experiment with one or more comparisons with the control falsely declared to be significant is specified at α.

EXAMPLE 9.10

Refer to the data of Example 9.3. Compare the two biological treatments and two chemical treatments to the control treatment using $\alpha = .05$.

Solution We want to determine whether the biological and chemical treatments have increased hay production, so we will conduct one-sided comparisons with the control.

1. From Example 9.3, we had $s_W^2 = .0153$ with df = 25 and $t = 5$ treatments including the control treatment. The critical value of the Dunnett procedure is found in the one-sided portion of Table 11 in the Appendix with

$$\alpha = .05 \qquad k = 5 - 1 = 4 \qquad v = 25$$

yielding $d_{.05}(4, 25) = 2.28$. Since $n_c = n_2 = n_3 = n_4 = n_5 = 6$, we have

$$D = d_\alpha(k, v)\sqrt{\frac{2s_W^2}{n}} = 2.28\sqrt{\frac{2(.0153)}{6}} = .163$$

2. We declare treatment mean μ_i greater than the control mean μ_c if $(\bar{y}_i - \bar{y}_c) \geq .163$. We can summarize the comparisons as shown in Table 9.8.

TABLE 9.8

Treatment	$(\bar{y}_i - \bar{y}_c)$	Comparison	Conclusion
Bio1	$(1.293 - 1.175) = .118$	$< D$	Not greater than control
Bio2	$(1.328 - 1.175) = .153$	$< D$	Not greater than control
Chm1	$(1.415 - 1.175) = .240$	$> D$	Greater than control
Chm2	$(1.500 - 1.175) = .325$	$> D$	Greater than control

3. We conclude that using either of the biological agents would result in an average hay production not greater than the production obtained using no agent on the fields. Thus, at the $\alpha = .05$ level, the biological agents are not effective in controlling weeds in the hay fields. However, the average hay production using the chemical agents appears to be greater than the hay production on fields with no weed agents. ■

When the sample sizes are not equal, the Dunnett procedure does not produce an experimentwise error rate equal to α. As noted earlier, Dunnett (1964) provided adjustments to the values given in Table 11 in the Appendix for the unequal sample sizes.

9.7 A Nonparametric Multiple-Comparison Procedure

The multiple-comparison procedures—Tukey's W, Dunnett, and Scheffé's S—all are based on the condition that the data are random samples from normal distributions with equal variances. In a number of situations (for example, income, percentage, or survival data), the normality condition is not valid, or the sample sizes are so small that it is not possible to conduct the diagnostics to verify the normality of the data. In a number of experiments, the recorded data are measured using an ordinal scale, and, hence, the relative ranks are the only meaningful measure, not the actual recorded measurements (for example, consumer rankings of products or tasters of new food products). In these types of situations, it is necessary to apply a procedure similar to the Wilcoxon rank sum test that is based on the ranks of the data. We will now describe a multiple-comparison procedure that is applicable when the data are not normally distributed.

The following procedure requires only that n_1 observations be randomly selected from population 1, n_2 observations from population 2, . . . , and n_t observations from population t. The t populations are identical except for possible differences in a shift parameter τ_i. Figure 8.9 demonstrates the type of situation in which this procedure would be applicable. We wish to determine which pairs of populations have a difference in their shift parameters—that is, have τ_i different from τ_j. For the multiple-comparison procedures in the previous sections, these were the same conditions with the exception that we imposed the additional condition that all t populations have a normal distribution. In that case, τ_i equals μ_i. This is not necessarily true for nonnormal distributions.

A Kruskal–Wallis–based nonparametric multiple-comparison procedure is summarized here.

Kruskal–Wallis Nonparametric Procedure:

1. Perform a Kruskal–Wallis test of H_0: $\tau_1 = \tau_2 = \cdots = \tau_t$ versus the alternative hypothesis that at least one of the τ_is differs from the rest.
2. If there is insufficient evidence to reject H_0 using the Kruskal–Wallis test, declare there is not sufficient evidence to determine a difference in the t populations and proceed no further.
3. If H_0 is rejected, calculate the $t(t-1)/2$ absolute differences $|\overline{R}_i - \overline{R}_j|$ for $i < j$, where $\overline{R}_i$ denotes the mean of the ranks for the measurements in sample i after the measurements from all t samples have been combined and then ranked from smallest to largest measurement.
4. Two populations are declared different if

$$|\overline{R}_i - \overline{R}_j| \geq KW_{ij}$$

where

$$\text{KW}_{ij} = \sqrt{h_\alpha\left(\frac{n_T(n_T+1)}{12}\right)\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

where $n_T = \sum_{i=1}^{t} n_i$ and h_α is the critical value for the Kruskal–Wallis test [Table A.12 in Hollander and Wolfe (1999)].

5. As an alternative when the n_js are large, we can approximate the critical value with

$$\text{KW}_{ij} \approx \frac{q_\alpha(t,\infty)}{\sqrt{2}}\sqrt{\frac{n_T(n_T+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

where $q_\alpha(t,\infty)$ is the critical value of the Studentized range from Table 10 in the Appendix.

6. The error rate that is controlled is an experimentwise error rate.

We will illustrate the application of the above procedure in the following example.

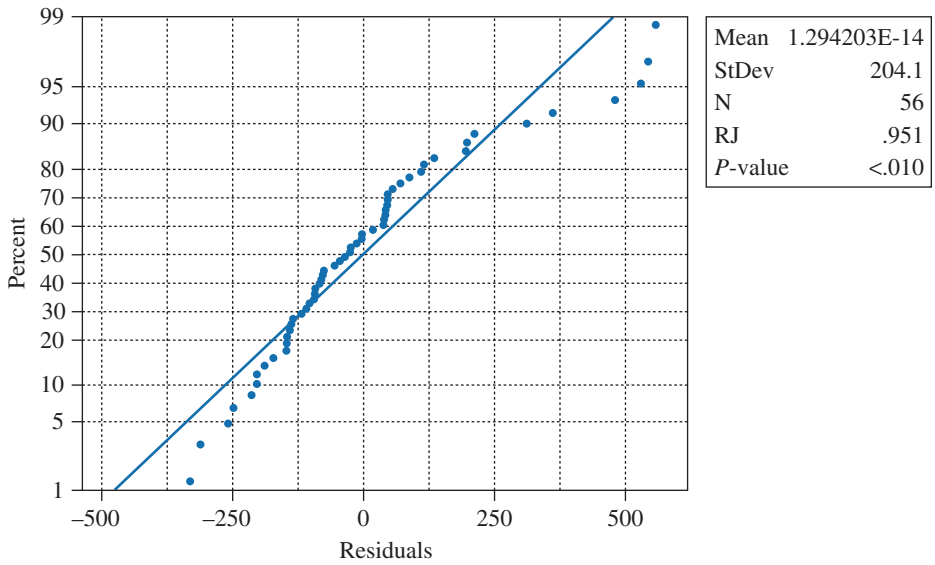
EXAMPLE 9.11

Of air pollutant gases, nitrogen dioxide is the most often encountered oxidant. Scientists have determined that nitrogen dioxide causes pathological alterations in the lung consistent with the diagnosis of emphysema. The researchers examined the protective power of a number of enzyme-inducing agents against the action of nitrogen dioxide on enzymes in the lung. A portion of that study will be described here. Fifty-six rats were randomly assigned to one of four treatment groups: control, 3-Methylcholanthrene (3-MC), allylisopropylacetamide (AIA), and p-aminobenzoic acid (PABA). In each experiment, the control and treatment animals were simultaneously exposed to nitrogen dioxide. The survival time (minutes)—that is, the time from the start of exposure to nitrogen dioxide until death—was determined. These values are given in Table 9.9.

TABLE 9.9
Survival times (minutes) of rats under four treatments

Subject	Control	3-MC	AIA	PABA
1	70.212	410.808	97.137	5.710
2	261.467	341.398	11.972	154.340
3	6.013	56.339	256.635	105.027
4	115.512	117.633	350.595	0.071
5	13.735	194.180	202.081	146.306
6	96.191	562.024	1.038	225.570
7	66.245	925.114	69.371	155.321
8	17.058	910.929	27.086	63.497
9	349.469	37.065	253.724	14.459
10	125.510	272.684	746.738	30.978
11	148.526	108.371	75.278	472.233
12	221.586	162.487	232.193	33.288
13	463.236	847.685	427.775	15.273
14	206.578	218.904	303.216	150.674

FIGURE 9.2
Normal probability plot of residuals



The researchers wanted to determine if the three treatments increased the survival times of the rats. A residual analysis of the above data yielded the normal probability plot in Figure 9.2.

The data deviate significantly from a normal distribution. Thus, the Kruskal–Wallis nonparametric procedure will be used to determine if any differences exist in the four treatments. The ranks of the data in the combined data set are given in Table 9.10.

TABLE 9.10
Ranks of the survival times

Subject	Control	3-MC	AIA	PABA
1	18	48	21	3
2	42	45	5	30
3	4	14	41	22
4	24	25	47	1
5	6	33	34	27
6	20	52	2	38
7	16	56	17	31
8	9	55	10	15
9	46	13	40	7
10	26	43	53	11
11	28	23	19	51
12	37	32	39	12
13	50	54	49	8
14	35	36	44	29
Mean	25.8	37.8	30.1	20.4

The computed value of the Kruskal–Wallis statistic was $H = 8.55$ with a p-value $= .036$. Thus, there was significant evidence of a difference in the distribution of survival times for the four treatments. Next, we will compare the six pairs of treatments to determine which pairs have significantly different shifts.

Because the sample sizes are relatively large, we will use the approximated method for computing the critical value for the multiple comparison:

$$\text{KW} \approx \frac{q_\alpha(t, \infty)}{\sqrt{2}} \sqrt{\frac{n_T(n_T + 1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

where $q_\alpha(t, \infty) = q_{.05}(4, \infty) = 3.63$, $n_1 = n_2 = n_3 = n_4 = 14$, and $n_T = 4(14) = 56$. Therefore, the critical value for all six comparisons is

$$\text{KW} \approx \frac{3.63}{\sqrt{2}} \sqrt{\frac{56(56 + 1)}{12}\left(\frac{1}{14} + \frac{1}{14}\right)} = 15.82$$

Thus, any pair of treatments having $|\overline{R}_i - \overline{Rj}| \geq 4.06$ will be declared significantly different. The results of the six comparisons are summarized in Table 9.11.

TABLE 9.11 Summary of the nonparametric multiple comparison

Treatment Pair	$\lvert R_i - R_j \rvert$	Conclusion
Control vs 3-MC	$\lvert 25.8 - 37.8 \rvert = 12$	Not significantly different
Control vs AIA	$\lvert 25.8 - 30.1 \rvert = 4.3$	Not significantly different
Control vs PABA	$\lvert 25.8 - 20.4 \rvert = 5.4$	Not significantly different
3-MC vs AIA	$\lvert 37.8 - 30.1 \rvert = 7.7$	Not significantly different
3-MC vs PABA	$\lvert 37.8 - 20.4 \rvert = 17.4$	Significantly different
AIA vs PABA	$\lvert 30.1 - 20.4 \rvert = 9.7$	Not significantly different

Thus, only one pair of treatments, 3-MC vs PABA, had significantly different survival times. ■

9.8 RESEARCH STUDY: Are Interviewers' Decisions Affected by Different Handicap Types?

There are approximately 50 million people in the United States who report having a handicap. Furthermore, it is estimated that the unemployment rate of noninstitutionalized handicapped people between the ages of 18 and 64 is nearly double the unemployment rate of people with no impairment. Thus, it appears that people with disabilities have a more difficult time obtaining employment. One of the problems confronting people having a handicap may be a bias by employers during the employment interview.

Defining the Problem

The paper "Interviewers' Decisions Related to Applicant Handicap Type and Rater Empathy" ***(Cesare et al., 1990)***, describes a study that examines these issues. The purposes of the study were to investigate whether different types of physical handicaps produce different levels of empathy in raters and to examine if interviewers' evaluations are affected by the type of handicap of the person being interviewed.

Five simulated employment interviews were videotaped. In order to minimize bias across videotapes, the same male actors (job applicant and interviewer) were used. Also, the same interview script, consisting of nine questions, was used in all five videotapes. The videotapes differed with respect to the type of applicant disability, all of which were depicted as being permanent disabilities. The five conditions were as follows: wheelchair, Canadian crutches, hard of hearing, leg amputee, and nonhandicapped (control).

Collecting the Data

A group of undergraduate students was randomly assigned to one of five experimental conditions that simulated an employment interview with an applicant having one of five conditions: used a wheelchair, used Canadian crutches, was hard of hearing, had a leg amputated, or was nonhandicapped (control). Each participant in the study was asked to rate the applicant's qualifications for a computer sales position based on the questions asked during the videotaped interview. Prior to viewing the videotape, each participant completed the Hogan Empathy Scale. The researchers decided to have each participant view only one of the five videotapes. Based on the variability in scores of raters in previous studies, the researchers decided they would require 14 raters for each videotape in order to obtain a precise estimate of the mean rating for each of the five handicap conditions. Seventy undergraduate students were selected to participate in the study, and 14 of them were randomly assigned to view each of the videotapes. After viewing the videotape, each participant rated the applicant on two scales: one an 11-item scale assessing the rater's liking of the applicant and a second 10-item scale that assessed the rater's evaluation of the applicant's job qualifications. For each scale, the average of the individual items form an overall assessment of the applicant. The researchers used these two variables to determine if different types of physical handicaps are reacted to differently by raters and to determine the effect of rater empathy on evaluations of handicapped applicants.

Some of the questions that the researchers were interested in included the following:

1. Is there a difference in the average empathy scores of the 70 raters?
2. Do the raters' average qualification scores differ across the five handicap conditions?
3. Which pairs of handicap conditions produced different average qualification scores?
4. Is the average rating for the control group (no handicap) greater than the average ratings for all types of handicapped applicants?
5. Is the average qualification rating for the hard-of-hearing applicant different from the average ratings for those applicants that had a mobility handicap.
6. Is the average qualification rating for the "crutches" applicant different from the average rating of the applicant who was either an amputee or in a wheelchair.
7. Is the average rating for the amputee applicant different from the average rating of the wheelchair applicant.

Summarizing the Data

The researchers conducted the experiments and obtained the following data from the 70 raters of the applicants. The data in Table 9.12 are a summary of the empathy values. The data in Table 9.13 are the applicant qualification scores of the 70 raters for the five handicap conditions.

TABLE 9.12
Empathy values across the five handicap conditions

Condition	Control (None)	Hard of Hearing	Canadian Crutches	One-Leg Amputee	Wheelchair
Mean	21.43	22.71	20.43	20.86	19.86
St. Dev.	3.032	3.268	3.589	3.035	3.348

TABLE 9.13
Ratings of applicant qualification across the five handicap conditions

Control	Hard of Hearing	Amputee	Crutches	Wheelchair
6.1	2.1	4.1	6.7	3.0
4.6	4.8	6.1	6.7	3.9
7.7	3.7	5.9	6.5	7.9
4.2	3.5	5.0	4.6	3.0
6.1	2.2	6.1	7.2	3.5
2.9	3.4	5.7	2.9	8.1
4.6	5.5	1.1	5.2	6.4
5.4	5.2	4.0	3.5	6.4
4.1	6.8	4.7	5.2	5.8
6.4	0.4	3.0	6.6	4.6
4.0	5.8	6.6	6.9	5.8
7.2	4.5	3.2	6.1	5.5
2.4	7.0	4.5	5.9	5.0
2.9	1.8	2.1	8.8	6.2

(The above data were simulated using the summary statistics of the ratings given in the paper.) A descriptive summary of these data is shown in Table 9.14.

TABLE 9.14 Descriptive statistics for ratings

Descriptive Statistics for Case Study

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Control	14	4.900	4.600	4.875	1.638	0.438
Hard of Hearing	14	4.050	4.100	4.108	1.961	0.524
Amputee	14	4.436	4.600	4.533	1.637	0.437
Crutches	14	5.914	6.300	5.925	1.537	0.411
Wheelchair	14	5.364	5.650	5.333	1.633	0.436

Variable	Minimum	Maximum	Q1	Q3
Control	2.400	7.700	3.725	6.175
Hard of Hearing	0.400	7.000	2.175	5.575
Amputee	1.100	6.600	3.150	5.950
Crutches	2.900	8.800	5.050	6.750
Wheelchair	3.000	8.100	3.800	6.400

The qualification scores were plotted in Figure 9.1. The boxplots display somewhat higher qualification scores from the raters viewing the crutches condition. The mean qualification scores for the hard-of-hearing and amputee conditions were somewhat smaller than those of the control and wheelchair conditions. The variabilities of the qualification scores were nearly the same for all five conditions.

Analyzing the Data

The objective of the study was to investigate whether an interviewer's evaluation of applicants for a job is affected by the physical handicap of the person being interviewed. Prior to testing hypotheses and making comparisons among the five treatments, we need to verify that the conditions required for the tests and multiple-comparison procedures to be valid have been satisfied in this study.

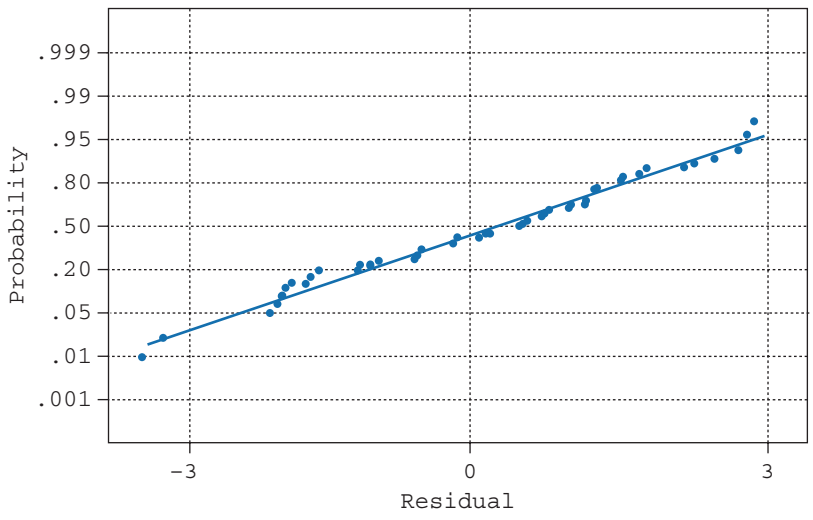
We observed in Figure 9.1 that the boxplots were of nearly the same width with no outliers and with whiskers of nearly the same length. The means and medians for the five groups of applicants were similar in size. Thus, the assumptions of AOV would appear to be satisfied. To confirm this observation, we computed the residuals and plotted them in a normal probability plot (see Figure 9.3).

From this plot, we can observe that, with the exception of two data values, the points fall nearly on a straight line. Also, the p-value for the test of the null hypothesis that the data have a normal distribution is .387. Thus, there is a strong confirmation that the five populations of ratings of applicants' qualifications have normal distributions.

Next, we can check on the equal variance assumption. From the summary statistics given in Table 9.14, we note that the standard deviations ranged from 1.537 to 1.961. Thus, there is very little difference in the sample standard deviations. To confirm this observation, we conduct a test of homogeneity of variance using the BFL test. We are testing the following:

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 \quad \text{versus} \quad H_a: \text{Variances are not all equal.}$$

FIGURE 9.3
Normal probability plot of residuals



We compute a value of L = .405. The critical value is $F_{.05,4,25} = 2.76$. Thus, we fail to reject H_0. Furthermore, we compute the p-value to be p-value $= P(F_{4,25} \geq .405) = .803$. Thus, we are confident that the condition of homogeneity of variance has not been violated in this study.

The condition of independence of the data would be checked by discussing with the researchers the manner in which the study was conducted. It would be important to make sure that the conditions in the room where the interview tape was viewed remained constant throughout the study so as to not introduce any distractions that could affect the raters' evaluations. Also, the initial check that the empathy scores were evenly distributed over the five groups of raters assures us a difference in empathy levels did not exist in the five groups of raters prior to their evaluation of the applicants' qualifications.

The research hypothesis is that the mean qualification ratings, μ_is, differ over the five handicap conditions:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a: At least one of the means differs from the rest.

The computer output for the AOV table is given here. The following notation is used in the output: control (C), hard of hearing (H), amputee (A), crutches (R), and wheelchair (W).

```
The GLM Procedure

ANOVA TABLE FOR COMPARING AVERAGE RATINGS OVER 5 TYPES OF HANDICAPS

Dependent Variable: RATING
                                      Sum of
Source                    DF         Squares    Mean Square  F Value  Pr > F
Model                      4     30.4780000      7.6195000     2.68  0.0394
Error                     65    185.0564286      2.8470220
Corrected Total           69    215.5344286


Dunnett's One-tailed t Tests for RATING

NOTE: This test controls the Type I experimentwise error for
comparisons of all treatments against a control.

Alpha                                          0.05
Error Degrees of Freedom                         65
Error Mean Square                          2.847022
Critical Value of Dunnett's t               2.20298
Minimum Significant Difference               1.4049

Comparisons significant at the 0.05 level are indicated by ***.

                 Difference
    HC              Between      Simultaneous 95%
Comparison            Means      Confidence  Limits

  R - C             1.0143      -Infinity   2.4192
  W - C             0.4643      -Infinity   1.8692
  A - C            -0.4643      -Infinity   0.9407
  H - C            -0.8500      -Infinity   0.5549
```

```
Tukey's Studentized Range (HSD) Test for RATING

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                                      0.05
Error Degrees of Freedom                     65
Error Mean Square                      2.847022
Critical Value of Studentized Range    3.96804
Minimum Significant Difference          1.7894

Means with the same letter are not significantly different.

Tukey
Grouping
                  Mean     N    HC

    A          5.9143    14    R
B   A          5.3643    14    W
B   A          4.9000    14    C
B   A          4.4357    14    A
B              4.0500    14    H

Dependent Variable: RATING

Contrast                DF    Contrast SS   Mean Square  F Value   Pr > F

Control vs. Handicap     1     0.01889286    0.01889286     0.01   0.9353
Hearing vs. Mobility     1    14.82148810   14.82148810     5.21   0.0258
Crutches vs. Amp.& Wheel 1     9.60190476    9.60190476     3.37   0.0709
```

From the output, we see that the p-value for the F test is .0394. Thus, there is a significant difference in the mean ratings across the five types of handicaps. We next investigate what types of differences exist in the ratings for the groups. We make a comparison of the control (C) group to the four groups having handicaps—crutches (R), wheelchair (W), amputee (A), and hard of hearing (H)—using the Dunnett procedure at the $\alpha_E = .05$ level. We use a one-sided test of whether any of the four handicap groups had a lower mean rating than did the control group:

$$H_0\colon\ \mu_i \geq \mu_C$$
$$H_a\colon\ \mu_i < \mu_C$$

We reach the conclusion that the mean rating for the control (no handicap) group is not significantly greater than the mean rating for any of the handicap groups. Next, we run a multiple procedure to determine which group pairs produced different mean ratings. The analysis uses the Tukey procedure with $\alpha = .05$, with the results displayed in the computer output. All handicap types with the same Tukey grouping letter have mean ratings that are not significantly different from each other. Thus, the mean rating from the applicant using crutches was significantly higher than the mean rating for the applicant who was hard of hearing. No other pairs were found to be significantly different. To investigate the size of the differences in the pairs of rating means for the five handicap conditions, we computed simultaneous 95% confidence intervals for the ten pairs of mean differences using the Tukey procedure. The intervals are provided in the following computer output.

```
Tukey's Studentized Range (HSD) Test for RATING

NOTE: This test controls the Type I experimentwise error rate.

Alpha                                      0.05
Error Degrees of Freedom                     65
Error Mean Square                      2.847022
Critical Value of Studentized Range    3.96804
Minimum Significant Difference          1.7894

Comparisons significant at the 0.05 level are indicated by ***.

           Difference
    HC        Between      Simultaneous 95%
Comparison      Means     Confidence Limits

  R - W        0.5500      -1.2394   2.3394
  R - C        1.0143      -0.7751   2.8037
  R - A        1.4786      -0.3108   3.2680
  R - H        1.8643       0.0749   3.6537  ***
  W - C        0.4643      -1.3251   2.2537
  W - A        0.9286      -0.8608   2.7180
  W - H        1.3143      -0.4751   3.1037
  C - A        0.4643      -1.3251   2.2537
  C - H        0.8500      -0.9394   2.6394
  A - H        0.3857      -1.4037   2.1751
```

Finally, several contrasts were constructed to evaluate the remaining questions posed by researchers. The questions along with the corresponding contrasts are given in Table 9.15.

TABLE 9.15

Question	Contrast
Control ratings vs. Handicap ratings	$4\mu_C - \mu_R - \mu_W - \mu_A - \mu_H$
Hearing ratings vs. Mobility handicap ratings	$0\mu_C - \mu_R - \mu_W - \mu_A + 3\mu_H$
Crutches ratings vs. Amputee wheelchair ratings	$0\mu_C + 2\mu_R - \mu_W - \mu_A + 0\mu_H$

From the computer output, we have p-values of .9353, .0258, and .0709 for testing the hypotheses:

$$H_0: l = 0 \quad \text{versus} \quad H_a: l \neq 0$$

We can use a Bonferroni procedure with $\alpha_E = .05$ to test the three sets of hypotheses. The individual comparison rate is set at $\alpha_I = \alpha_E/3 = .05/3 = .0167$. Thus, if the p-value for any one of the three F tests of the significance of the contrasts is less than .0167, we will declare that contrast to be significantly different from 0. Because the three p-values were .9353, .0258, and .0709, none of the three contrasts is significantly different from 0.

The only significant difference found in the five mean ratings was between the applicant with a hearing handicap and the applicant using crutches. The researchers discussed in detail in the article why this difference may have occurred.

Reporting the Conclusions

We would need to write a report summarizing our findings of this study. We would need to include to following:

1. Statement of objective for study
2. Description of study design, how raters were selected, and how interviews were conducted

3. Discussion of the generalizability of results from the study
4. Numerical and graphical summaries of data sets
5. Description of all inference methodologies
   - AOV table and F test
   - multiple-comparison procedures, contrasts, and confidence intervals
   - verification that all necessary conditions for using inference techniques were satisfied
6. Discussion of results and conclusions
7. Interpretation of findings relative to previous studies
8. Recommendations for future studies
9. Listing of data sets

9.9 Summary and Key Formulas

We presented three multiple-comparison procedures (Bonferroni t test, Tukey's W, and Scheffé's S for making pairwise comparisons of t population means. Each of these procedures controls the experimentwise error rate. There are numerous other procedures, such as Fisher's LSD and Newman–Kuels, that are more powerful; that is, these procedures will tend to declare more pairs of means to be different. However, these procedures do not control the experimentwise error rate, which results in uncertainty about the probability of Type I errors when using these procedures. For this reason, many statisticians would not recommend using either of these procedures.

A comparison of the three procedures discussed in this chapter can be made by considering the magnitude of the difference in the sample means, $|\bar{y}_i - \bar{y}_k|$, needed to declare the population means, μ_i and μ_k, to be different. The larger the magnitude of the difference, the more conservative the procedure—that is, the less likely it is to declare a pair of population means to be different. To illustrate these comparisons, we will use the data from the five populations in Example 9.3.

In computing the critical magnitude of $|\bar{y}_i - \bar{y}_k|$ for the Bonferroni t test, we are considering 10 pairs of means. This would require using the upper $\frac{\alpha/2}{10}$ percentile from the t distribution with df $= n - t = 25$. Thus, the critical values for an $\alpha = .05$ Bonferroni t test would be

$$|\bar{y}_i - \bar{y}_k| \geq t_{.0025,25}\sqrt{s_W^2\left(\frac{1}{n_i} + \frac{1}{n_k}\right)} = 3.0782\sqrt{(.0153)\left(\frac{1}{6} + \frac{1}{6}\right)} = .2198$$

In computing the critical magnitude of $|\bar{y}_i - \bar{y}_k|$ for the Scheffé's S, we are considering the differences in 10 pairs of means, which can be represented by the contrasts $l = \mu_i = \mu_k$. Thus, the five coefficients in the contrasts are three 0s, $+1$, and -1. This leads to a critical value for the Scheffé's S of

$$|\bar{y}_i - \bar{y}_k| \geq \sqrt{s_W^2\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}\sqrt{(t-1)F_{\alpha,\,t-1,\,n-t}}$$

$$= \sqrt{(.0153)\left(\frac{1}{6} + \frac{1}{6}\right)}\sqrt{(5-1)(2.7587)} = .2372$$

The critical value for the Tukey's W was computed in Example 9.8 to be

$$|\bar{y}_i - \bar{y}_k| \geq .2097$$

The value of the Tukey's W is smaller than the value for the Bonferroni t test, which is smaller than the value for the Scheffé's S. This will result in the

Tukey's W procedure declaring as many or more pairs of means to be different than the Bonferroni t test and the Scheffé's S procedure. The Tukey's W procedure is the least conservative of the three procedures, while maintaining the specified level of experimentwise error rate, and would be the procedure used in most situations.

In those situations where the data are not from a normally distributed population, we presented a distribution-free procedure based on the Kruskal–Wallis statistics. Thus, when encountering data that are measured on an ordinal scale, we do not need to compromise our normal-based procedure but can apply a procedure specifically designed for data based solely on their relative ranks.

Key Formulas

1. Tukey's W procedure

$$W = q_\alpha(t, \nu)\sqrt{\frac{s_W^2}{n}}$$

2. Tukey–Kramer W^* procedure

$$W^* = \frac{q_\alpha(t, \nu)}{\sqrt{2}}\sqrt{s_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

3. Dunnett's procedure

$$D = d_\alpha(k, \nu)\sqrt{\frac{2s_W^2}{n}}$$

4. Scheffé's S method

$$S = \sqrt{\hat{V}(\hat{l})}\sqrt{(t-1)F_{\alpha, \mathrm{df}_1, \mathrm{df}_2}}$$

where

$$\hat{V}(\hat{l}) = s_W^2\sum\frac{a_i^2}{n_i}$$

5. Nonparametric procedure

$$\mathrm{KW}_{ij} \approx \frac{q_\alpha(t, \infty)}{\sqrt{2}}\sqrt{\frac{n_T(n_T+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

9.10 Exercises

9.1 Introduction

Soc. **9.1** In the research study concerning interviewers' decisions:

a. What are the populations of interest?

b. What are some of the limitations of this study based on the participating subjects?

Soc. **9.2** In the research study concerning interviewers' decisions:

a. Describe how the subjects in this experiment could have been selected so as to satisfy the randomization requirements?

b. State several research hypotheses, other than those given in the abstract, that may have been of interest to the researchers.

9.2 Linear Contrasts

Basic **9.3** In an experiment with $t = 4$ populations means, consider the four linear combinations of those means.

$$l_1 = \mu_1 - 3\mu_2 + \mu_3 + \mu_4$$
$$l_2 = \mu_1 + \mu_2 - 2\mu_4$$
$$l_3 = \mu_1 + \mu_2 + \mu_3 + \mu_4$$
$$l_4 = \mu_1 + \mu_2 - 3\mu_3 + \mu_4$$

a. Which of the four linear combinations are contrasts?

b. Which pairs of *contrasts* are orthogonal?

c. Suppose we have two contrasts:

$$l_1 = \mu_1 + \mu_2 + \mu_3 - 3\mu_4 \quad \text{and} \quad l_2 = \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \mu_4$$

Is testing $H_0 : l_1 = 0$ equivalent to testing $H_0 : l_2 = 0$? Justify your answer.

Basic **9.4** In an experiment with $t = 4$ population means and sample sizes of $n_1 = 5$, $n_2 = 6$, $n_3 = 4$, and $n_4 = 8$, consider the four linear combinations of the sample means:

$$\hat{l}_1 = \bar{y}_1 - 3\bar{y}_2 + \bar{y}_3 + \bar{y}_4$$
$$\hat{l}_2 = \bar{y}_1 + \bar{y}_2 - 2\bar{y}_4$$
$$\hat{l}_3 = \bar{y}_1 + \bar{y}_2 + \bar{y}_3 + \bar{y}_4$$
$$\hat{l}_4 = \bar{y}_1 + \bar{y}_2 - 3\bar{y}_3 + \bar{y}_4$$

a. Which of the four linear combinations are contrasts?
b. Which pairs of *contrasts* are orthogonal?

Soc. **9.5** In the abstract to the research study described earlier in this chapter, the researchers were interested in answering several questions concerning the difference in the raters' reactions to various handicaps. For each of the following questions, write a contrast in the five condition mean ratings that would attempt to answer the researchers' question.

a. Question 1: Is the average rating for the control (no handicap) group greater than the average ratings for all types of handicapped applicants?
b. Question 2: Is the average qualification rating for the hard-of-hearing applicant different from the average ratings for those applicants that had a mobility handicap?
c. Question 3: Is the average qualification rating for the crutches applicant different from the average rating of the applicant who was either an amputee or in a wheelchair?
d. Question 4: Is the average rating for the amputee applicant different from the average rating of the wheelchair applicant?

Soc. **9.6** Refer to Exercise 9.5. For each pair of contrasts, determine if it is orthogonal:

a. Question 1 and Question 2
b. Question 1 and Question 3
c. Question 1 and Question 4
d. Question 2 and Question 3
e. Question 2 and Question 4
f. Question 3 and Question 4
g. Are the four contrasts mutually orthogonal?

Pol. Sci. **9.7** Refer to Example 8.6. The political action group was interested in determining regional differences in the public's opinion concerning air pollution. Write a contrast in the four population means to answer each of the following questions.

a. Question 1: Is the proportion of people who thought the EPA's standards are not stringent enough different for the people living in the East compared to the people living in the West?
b. Question 2: Is the proportion of people who thought the EPA's standards are not stringent enough different for the people living in the Northeast compared to the people living in the other three regions?
c. Question 3: Is the proportion of people who thought the EPA's standards are not stringent enough different for the people living in the Northeast compared to the people living in the Southeast?
d. Simultaneously test if the three contrasts are different from 0 using an $\alpha = .05$ test.
e. Are the three contrasts mutually orthogonal?

9.3 Which Error Rate Is Controlled?

Basic **9.8** In a study of 10 new producers of iron supplements, nine contrasts in the mean iron level in the supplements were constructed by the quality control department for comparing various characteristics of the producers.

a. In order to achieve an experimentwise error rate of .05, what value should be selected for the value of α_I?
b. What is the critical value for the F statistic for testing the nine contrasts if there were six samples of the supplement taken from each of the 10 producers?

Basic **9.9** In a study comparing the mean yield of nine formulations of a fertilizer, the researcher constructed eight contrasts for comparing various aspects of the nine formulations. The researcher had selected a value of .005 for α_I in conducting each of the eight tests. Place an upper bound on the experimentwise error rate for the eight tests.

Basic **9.10** In Exercise 9.8, the Bonferroni procedure was used to ensure an experimentwise error rate of .05; that is, the probability of one or more Type I errors in conducting the nine tests is at most .05. The Bonferroni procedure is labeled as a conservative procedure because the actual experimentwise error rate is most likely to be somewhat less than .05. This is a positive aspect of the procedure in that the chance of Type I errors is even less than specified. The old adage "There is no free lunch" applies in this situation. State some of the negative aspects of using the Bonferroni procedure.

Supplementary Exercises

Soc. **9.11** Refer to Exercise 3.55.

- **a.** Is the average of the mean expenditures of families with three or fewer members less than the average of the mean expenditures for families with four or more members? Use $\alpha = .05$.
- **b.** Which pairs of the five groups have significantly different mean expenditures? Use an experimentwise error rate of .05.

Bio. **9.12** Refer to Exercise 7.18. The wildlife biologist was interested in determining if the mean weight of deer raised in a zoo would be lower than that of deer raised in a more uncontrolled environment—for example, raised either in the wild or on a ranch.

- **a.** Use a multiple-comparison procedure to determine if the mean weight of the deer raised in the wild or on a ranch is significantly higher than the mean weight of deer raised in a zoo.
- **b.** Write a linear contrast to compare the average weight of deer raised in a zoo or on a ranch to the mean weight of deer raised in the wild.
- **c.** Test at the $\alpha = .05$ level whether your contrast in part (b) is significantly different from zero. What conclusions can you draw from this test?

Med. **9.13** Researchers conducted an experiment to compare the effectiveness of four new weight-reducing agents to that of an existing agent. The researchers randomly divided a random sample of 50 males into five equal groups with preparation A1 assigned to the first group, A2 to the second group, and so on. They then gave a prestudy physical to each person in the experiment and told him how many pounds overweight he was. A comparison of the mean numbers of pounds overweight for the groups showed no significant differences. The researchers then began the study program, and each group took the prescribed preparation for a fixed period of time. The weight losses recorded at the end of the study period are given here:

A_1	12.4	10.7	11.9	11.0	12.4	12.3	13.0	12.5	11.2	13.1
A_2	9.1	11.5	11.3	9.7	13.2	10.7	10.6	11.3	11.1	11.7
A_3	8.5	11.6	10.2	10.9	9.0	9.6	9.9	11.3	10.5	11.2
A_4	12.7	13.2	11.8	11.9	12.2	11.2	13.7	11.8	12.2	11.7
S	8.7	9.3	8.2	8.3	9.0	9.4	9.2	12.2	8.5	9.9

The standard (existing) agent is labeled agent S, and the four new agents are labeled A_1, A_2, A_3, and A_4. Run an analysis of variance to determine whether there are any significant differences among the five weight-reducing agents. Use $\alpha = .05$. Do any of the AOV assumptions appear to be violated? What conclusions do you reach concerning the mean weight loss achieved using the five different agents?

9.14 Refer to Exercise 9.13. Determine the significantly different pairs of means using the Tukey's W with $\alpha = .05$.

Med. **9.15** Refer to Exercises 9.13 and 9.14.

- **a.** Use a Bonferroni t test to determine which pairs of means are significantly different. Use $\alpha_E = .05$.
- **b.** Use Scheffé's S procedure to determine which pairs of means are significantly different. Use $\alpha_E = .05$.
- **c.** Which of the three procedures determined the largest number of significantly different pairs of means? The fewest?

9.16 Refer to Exercise 9.13. The researcher wants to determine which of the new agents produced a significantly larger mean weight loss in comparison to the standard agent. Use $\alpha = .05$ in making this determination.

9.17 Refer to Exercise 9.13. Suppose the new weight-loss agents were of the following form:

A_1: Drug therapy with exercise and counseling
A_2: Drug therapy with exercise but no counseling
A_3: Drug therapy with counseling but no exercise
A_4: Drug therapy with no exercise and no counseling

Construct contrasts to make comparisons among the agent means that will address the following:

- **a.** Compare the mean for the standard agent to the average of the means for the four new agents.
- **b.** Compare the mean for the agents with counseling to those without counseling. (Ignore the standard.)
- **c.** Compare the mean for the agents with exercise to those without exercise. (Ignore the standard.)
- **d.** Compare the mean for the agents with counseling to the standard.

9.18 Refer to Exercise 9.17. Use a multiple-testing procedure to determine at the $\alpha = .05$ level which of the contrasts is significantly different from zero. Interpret your findings relative to the researchers' question about finding the most effective weight-loss method.

Ag.

9.19 Refer to Exercise 8.7.

- **a.** Did continuous grazing result in a greater mean soil density than the grazing regimens in which there was a no grazing period?
- **b.** How large a difference is there in the mean soil densities for the three grazing regimens?

9.20 Refer to Exercise 8.28.

- **a.** Compare the mean yields of herbicide 1 and herbicide 2 to the control treatment. Use $\alpha = .05$.
- **b.** Should the procedure you used in part (a) be a one-sided or a two-sided procedure?
- **c.** Interpret your findings in part (a).

9.21 Refer to Exercise 8.31.

- **a.** Compare the mean scores for the three divisions using an appropriate multiple-comparison procedure. Use $\alpha = .05$.
- **b.** What can you conclude about the differences in mean scores and the nature of the divisions from which any differences arise?

Ag.

9.22 **The** nitrogen contents of red clover plants inoculated with three strains of *Rhizobium* are given here:

3DOK1	3DOK5	3DOK7
19.4	18.2	20.7
32.6	24.6	21.0
27.0	25.5	20.5
32.1	19.4	18.8
33.0	21.7	18.6
	20.8	20.1
		21.3

- **a.** Is there evidence of a difference in the effects of the three treatments on the mean nitrogen content? Analyze the data completely, and draw conclusions based on your analysis. Use $\alpha = .01$.
- **b.** Was there any evidence of a violation in the conditions required to conduct your analysis in part (a)?

Vet.

9.23 Researchers conducted a study of the effects of three drugs on the fat content of the shoulder muscles in labrador retrievers. They divided 80 dogs at random into four treatment groups. The dogs in group A were the untreated controls, while groups B, C, and D received one of three new heartworm medications in their diets. Five dogs randomly selected from each of the four groups received treatment for periods varying from 4 months to 2 years. The percentage of fat content of the shoulder muscles was determined and is given here.

	Treatment Group			
Examination Time	**A**	**B**	**C**	**D**
4 months	2.84	2.43	1.95	3.21
	2.49	1.85	2.67	2.20
	2.50	2.42	2.23	2.32
	2.42	2.73	2.31	2.79
	2.61	2.07	2.53	2.94
8 months	2.23	2.83	2.32	2.45
	2.48	2.59	2.36	2.49
	2.48	2.53	2.46	2.95
	2.23	2.73	2.04	2.05
	2.65	2.26	2.30	2.31
1 year	2.30	2.70	2.85	2.53
	2.30	2.54	2.75	2.73
	2.38	2.70	2.62	2.65
	2.05	2.81	2.50	2.84
	2.13	2.70	2.69	2.92
2 years	2.64	3.24	2.90	2.91
	2.56	3.71	3.02	2.89
	2.30	2.95	3.78	3.21
	2.19	3.01	2.96	2.89
	2.45	3.08	2.87	2.68
Mean	2.411	2.694	2.605	2.698

Under the assumption that conditions for an AOV were met, the researchers then computed an AOV to evaluate the difference in mean percentages of fat content for dogs under the four treatments. The AOV computations did not takes into account the length of time on the medication. The AOV is given here.

Source	**df**	**SS**	**MS**	***F* ratio**	***p*-value**
Treatments	3	1.0796	.3599	3.03	.0345
Error	76	9.0372	.1189		
Totals	79	10.1168			

a. Is there a significant difference in the mean percentages of fat content in the four treatment groups? Use $\alpha = .05$.

b. Do any of the three treatments for heartworm appear to have increased the mean percentage of fat content over the level in the control group?

9.24 Refer to Exercise 9.23. Suppose the researchers conjectured that the new medications caused an increase in fat content and that this increase accumulated as the medication was continued in the dogs. How could we examine this question using the data given?

Med. **9.25** The article ***"The Ames Salmonell/Microsome Mutagenicity Assay: Issues of Inference and Validation" [Journal of American Statistical Association (1989) 84:651–661]*** discusses the importance of chemically induced mutation for human health and the biological basis for the primary in vitro assay for mutagenicity, the Ames Salmonell/microsome assay. In an Ames test, the response obtained from a single sample is the number of visible colonies that result from plating approximately 10^8 microbes. A common protocol for an Ames test includes multiple samples at a control dose and four or five logarithmically spaced doses of a test compound. The following data are from one such experiment with 20 samples per dose level. The dose levels were μg/sample.

Dose	Number of Visible Colonies																				$\bar{y}_{i.}$	s_i^2
Control	11	13	14	14	15	15	15	15	16	17	17	18	18	19	20	21	22	23	25	27	17.8	17.5
.3	39	39	42	43	44	45	46	50	50	50	51	52	52	52	55	61	62	63	67	70	51.7	81.0
1.0	88	90	92	92	102	104	104	106	109	113	117	117	119	119	120	120	121	122	130	133	110.9	175.4
3.0	222	233	251	251	253	255	259	275	276	283	284	294	299	301	306	312	315	323	337	340	283.5	1,131.5
10.0	562	587	595	604	623	666	689	692	701	702	703	706	710	714	733	739	763	782	786	789	692.3	4,584.4

We want to determine whether there is an increasing trend in the mean number of colonies as the dose level increases. One method of making such a determination is to use a contrast with constants a_i determined in the following fashion. Suppose the treatment levels are t values of a continuous variable x: $x_1, x_2, \ldots, x_t$. Let $a_i = x_i - \bar{x}$ and $\hat{l} = \sum a_i \bar{y}_i$. If $\hat{l}$ is significantly different from zero and positive, then we state there is a positive trend in the μ_is. If $\hat{l}$ is significantly different from zero and negative, then we state there is a negative trend in the μ_is. In this experiment, the dose levels are the treatments $x_1 = 0$, $x_2 = .3$, $x_3 = 1.0$, $x_4 = 3.0$, and $x_5 = 10.0$ with $\bar{x} = 2.86$. Thus, the coefficients for the contrasts are $a_1 = 0 - 2.86 = -2.86$, $a_2 = 0.3 - 2.86 = -2.56$, $a_3 = 1.0 - 2.86 = -1.86$, $a_4 = 3.0 - 2.86 = +.14$, and $a_5 = 10.0 - 2.86 = +7.14$. We therefore need to evaluate the significance of the following contrast in the treatment means given by $-2.86\bar{y}_c - 2.56\bar{y}_{.3} - 1.86\bar{y}_{1.0} + 0.14\bar{y}_{3.0} + 7.14\bar{y}_{10.0}$. If the contrast is significantly different from zero and is positive, we conclude that there is an increasing trend in the dose means.

a. Test whether there is an increasing trend in the dose mean. Use $\alpha = .05$.

b. Do there appear to be any violations in the conditions necessary to conduct the test in part (a)? If there are violations, suggest a method that would enable us to validly test whether the positive trend exists.

9.26 In the research study concerning the evaluation of interviewers' decisions related to applicant handicap type, the raters were 70 undergraduate students, and the same male actors, both job applicant and interviewer, were used in all the videotapes of the job interview.

a. Discuss the limitations of this study in regard to using the undergraduate students as the raters of the applicant's qualifications for the computer sales position.

b. Discuss the positive and negative points of using the same two actors for all five interview videotapes.

c. Discuss the limitations of not varying the type of job being sought by the applicant.

Med. **9.27** The paper ***"The Effect of an Endothelin-Receptor Antagonist, Bosentan, on Blood Pressure in Patients with Essential Hypertension" [The New England Journal of Medicine (1998)]*** discussed the contribution of bosentan to blood pressure regulation in patients with essential hypertension. The study involved 243 patients with mild-to-moderate essential hypertension. After a placebo run-in period, patients were randomly assigned to receive one of four oral doses of bosentan (100, 500, or 1,000 mg once daily or 1,000 mg twice daily) or a placebo. The blood pressure was measured before treatment began and after a 4-week treatment period. The primary end point of the study was the change in blood pressure from the baseline obtained prior to treatment to the blood pressure at the conclusion of the 4-week treatment period. A summary of the data is given in the following table.

	Blood Pressure Change				
	Placebo	**100 mg**	**500 mg**	**1,000 mg**	**2,000 mg**
Diastolic pressure					
Mean	−1.8	−2.5	−5.7	−3.9	−5.7
Standard deviation	6.71	7.30	6.71	7.21	7.30
Systolic pressure					
Mean	−0.9	−2.5	−8.4	−10.3	−10.3
Standard deviation	11.40	11.94	11.40	11.80	11.94
Sample size	45	44	45	43	44

a. Which of the dose levels were associated with a significantly greater reduction in the diastolic pressure in comparison to the placebo? Use $\alpha = .05$.
b. Why was it important to include a placebo treatment in the study?
c. Using just the four treatments (ignore the placebo), construct a contrast to test for an increasing linear trend in the size of the systolic pressure reductions as the dose levels are increased. See Exercise 9.25 for the method for creating such a contrast.
d. Use Tukey's W procedure to test for pairwise differences in the mean systolic blood pressure reductions for the four treatment doses. Use $\alpha = .05$.
e. The researchers referred to their study as a *double-blind* study. Explain the meaning of this terminology.

9.28 Refer to Exercise 8.23.
a. Use a nonparametric procedure to compare the mean reliability of the seven plants.
b. Even though the necessary conditions are not satisfied, use the Tukey's W procedure to group the seven nuclear power plants based on their mean reliability.
c. Compare your results in part (b) to the groupings obtained in part (a).

Engin. **9.29** Refer to Exercise 8.27.
a. Use a nonparametric procedure to group the suppliers based on their mean deviations. Use an experimentwise error rate of .05.
b. Use the Tukey's W procedure to group the suppliers based on their mean deviations. Use an experimentwise error rate of .05.
c. Compare the two sets of groupings. Why is the nonparametric procedure more appropriate in this situation?

Hort. **9.30** Refer to Exercise 8.29.
a. Compare the mean discoloration scores of groups II, III, and IV to the control group. Use an experimentwise error rate of .05.
b. Use the Tukey's W procedure to compare the mean discoloration scores of groups II, III, and IV. Use an experimentwise error rate of .05.
c. Are there any inconsistencies in your conclusions in parts (a) and (b)?

Hort. **9.31** Refer to Exercise 9.30.
a. Use a nonparametric procedure to compare the mean discoloration scores of groups II, III, and IV. Use an experimentwise error rate of .05.
b. Compare your results in part (a) to your conclusions from Exercise 9.30. Why is the Tukey's W procedure more appropriate in this situation?

CHAPTER 10

Categorical Data

10.1 Introduction and Abstract of Research Study

categorical or count data

Up to this point, we have been concerned primarily with sample data measured on a quantitative scale. However, we sometimes encounter situations in which levels of the variable of interest are identified by name or rank only and we are interested in the number of observations occurring at each level of the variable. Data obtained from these types of variables are called **categorical** or **count data.** For example, an item coming off an assembly line may be classified into one of three quality classes: acceptable, repairable, or reject. Similarly, a traffic study might require a count and classification of the type of transportation used by commuters along a major access road into a city. A pollution study might be concerned with the number of different alga species identified in samples from a lake and the number of times each species is identified. A consumer protection group might be interested in the results of a prescription fee survey to compare prices of some common medications in different areas of a large city.

In this chapter, we will examine specific inferences that can be made from experiments involving categorical data.

Abstract of Research Study: Does Gender Bias Exist in the Selection of Students for Vocational Education?

Although considerable progress has been made in recent years, barriers persist for women in education. The American Civil Liberties Union (ACLU) has at its website several articles that advance the notion that gender bias continues in the determination of career education where girls are generally found in programs that educate them for the traditionally female (and low-wage) fields of child care, cosmetology, and health assistance, whereas boys are found in higher proportions in courses preparing them for high-wage plumbing, welding, and electrician jobs. In some instances, this is the result of discriminatory steering by counselors and teachers, harassment by peers, and other forms of discrimination, which result from a failure to enforce governmental regulations and laws. The data support the contention that women still fall behind men in earning doctorates and professional degrees. While girls in high school are enrolled in nearly the same proportions as boys in high-level math and science courses, they are less likely to earn postsecondary degrees in these topics and are particularly grossly underrepresented in the fields of engineering and computer science. The June 2002 report ***Title IX at 30, Report Card on Gender Equity*** by the National Coalition for Women and Girls in Education reveals that female students are steered away from advanced computer courses and are often not informed of opportunities to take technology-related courses. Even in the area of athletics, where the most noticeable advancements for girls have occurred, male sports continue to receive more money than female sports at many colleges and universities.

These examples have been used to argue that there are continuing gender inequities in education. Determining whether these differences in educational opportunities for boys and girls are due to gender discrimination is both legally and morally important. However, it is very difficult to demonstrate that discrimination has occurred using just the enrollment data for students in various high school vocational programs. The data sets and summary figures that illustrate these important issues are given in the last section of this chapter. They will illustrate how aggregate data sets can often lead to misleading conclusions about important social issues.

10.2 Inferences About a Population Proportion π

In the binomial experiment discussed in Chapter 4, each trial results in one of two outcomes, which we labeled as either a success or a failure. We designated π as the probability of a success and $(1 - \pi)$ as the probability of a failure. Then the probability distribution for y, the number of successes in n identical trials, is

$$P(y) = \frac{n!}{y!(n-y)!}\pi^{y}(1-\pi)^{n-y}$$

The point estimate of the binomial parameter π is one that we would choose intuitively. In a random sample of n from a population in which the proportion of elements classified as successes is π, the best estimate of the parameter π is the

sample proportion of successes. Letting y denote the number of successes in the n sample trials, the sample proportion is

$$\hat{\pi} = \frac{y}{n}$$

We observed in Section 4.13 that y possesses a mound-shaped probability distribution that can be approximated by using a normal curve when

$$n \geq \frac{5}{\min(\pi, 1 - \pi)} \quad \text{(or, equivalently, } n\pi \geq 5 \text{ and } n(1 - \pi) \geq 5\text{)}$$

In a similar way, the distribution of $\hat{\pi} = y/n$ can be approximated by a normal distribution with a mean and a standard error as given here.

Mean and Standard Error of $\hat{\pi}$

$$\mu_{\hat{\pi}} = \pi$$

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

The normal approximation to the distribution of $\hat{\pi}$ can be applied under the same condition as that for approximating y by using a normal distribution. In fact, the approximation for both y and $\hat{\pi}$ becomes more precise for large n.

A confidence interval can be obtained for π using the methods of Chapter 5 for μ by replacing $\bar{y}$ with $\hat{\pi}$ and $\sigma_{\bar{y}}$ with $\sigma_{\hat{\pi}}$. A general $100(1 - \alpha)\%$ confidence interval for the binomial parameter is given here.

Confidence Interval for π with Confidence Coefficient of $(1 - \alpha)$

$$\hat{\pi} \pm z_{\alpha/2}\hat{\sigma}_{\hat{\pi}} \quad \text{or} \quad (\hat{\pi} - z_{\alpha/2}\hat{\sigma}_{\hat{\pi}}, \hat{\pi} + z_{\alpha/2}\hat{\sigma}_{\hat{\pi}})$$

where

$$\hat{\pi} = \frac{y}{n} \quad \text{and} \quad \hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

EXAMPLE 10.1

Researchers in the development of new treatments for cancer patients often evaluate the effectiveness of new therapies by reporting the proportion of patients who survive for a specified period of time after completion of the treatment. A new genetic treatment of 870 patients with a particular type of cancer resulted in 330 patients surviving at least 5 years after treatment. Estimate the proportion of all patients with the specified type of cancer who would survive at least 5 years after being administered this treatment. Use a 90% confidence interval.

Solution For these data,

$$\hat{\pi} = \frac{330}{870} = .38$$

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{(.38)(.62)}{870}} = .016$$

The confidence coefficient for our example is .90. Recall from Chapter 5 that we can obtain $z_{\alpha/2}$ by looking up the z-value in Table 1 in the Appendix corresponding to an area of $\alpha/2$. For a confidence coefficient of .90, the z-value corresponding to an area of .05 is 1.645. Hence, the 90% confidence interval on the proportion of cancer patients who will survive at least 5 years after receiving the new genetic treatment is

$$.38 \pm 1.645(.016) \quad \text{or} \quad .38 \pm .026 = (.354, .406) \quad \blacksquare$$

The confidence interval for π just presented is the standard confidence interval in most textbooks. It is often referred to as the Wald confidence interval. This confidence interval for π is based on a normal approximation to the binomial distribution. The rule that we specified in Chapter 4 was that both $n\pi$ and $n(1-\pi)$ should be at least 5. However, recent articles have shown that even when this rule holds, the Wald confidence interval may not be appropriate. When the sample size is too small and/or $\pi < .2$ or $\pi > .8$, the Wald confidence interval for π will often be quite inaccurate. That is, the true level of confidence can be considerably lower than the nominal level, or the confidence interval can be considerably wider than necessary for the nominal level of confidence. These articles discuss how slight adjustments to the Wald confidence interval can result in a considerable improvement in its performance.

The required adjustments to the traditional confidence interval for π involve moving $\hat{\pi}$ slightly away from 0 and 1. This adjustment was first introduced in a paper by Edwin Wilson in 1927 and involved a considerable amount of calculation. A recent modification to Wilson's confidence interval that performs nearly as well is contained in Agresti and Coull (1998). We will refer to this interval as the Wilson–Agresti–Coull (WAC) confidence interval. In the following, let y be the number of successes in n independent trials or the number of occurrences of an event in a random sample of n items selected from a large population.

WAC Confidence Interval for π with Confidence Coefficient of $(1-\alpha)$

Adjustments to y, n, and $\hat{\pi}$:

$$\tilde{y} = y + .5z_{\alpha/2}^2, \quad \tilde{n} = n + z_{\alpha/2}^2, \quad \tilde{\pi} = \frac{\tilde{y}}{\tilde{n}}$$

WAC Confidence Interval for π:

$$\tilde{\pi} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{\tilde{n}}} \quad \text{or} \quad \tilde{\pi} - z_{\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{\tilde{n}}},\ \tilde{\pi} + z_{\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{\tilde{n}}}$$

For a 95% confidence interval, the WAC interval is essentially *add 2 to y and 4 to n,* and then apply the standard Wald formula.

In the Agresti and Coull (1998) article, the authors state, "Our results suggest that (if one uses the WAC) interval, it is not necessary to present sample size rules $(n\pi > 5, n(1-\pi) > 5)$, since…[the WAC confidence interval] behaves adequately for practical application for essentially any n regardless of the value of π." In the article by Brown, Cai, and DasGupta (2001), the authors recommend using the WAC confidence interval whenever $n \geq 40$. When $n < 40$, the authors recommend the original Wilson confidence interval or a Bayesian-based procedure. However, they further comment that even for small sample sizes, the WAC

confidence interval is much preferable to the standard Wald procedure. The following example will illustrate the calculations involved in the WAC confidence interval.

EXAMPLE 10.2

The water department of a medium-sized city is concerned about how quickly its maintenance crews react to major breaks in the water lines. A random sample of 50 requests for repairs is analyzed, and 43 of the 50 requests were responded to within 24 hours. Construct a 95% confidence interval for the proportion π of requests for repair that are handled within 24 hours.

Solution Using the traditional method, the 95% confidence interval for π is computed as follows:

$$\hat{\pi} = \frac{43}{50} = .86 \quad \text{and} \quad \hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{.86(1 - .86)}{50}} = .0491$$

The confidence coefficient for this example is .95; therefore, the appropriate value for $z_{\alpha/2} = z_{.025} = 1.96$. Hence, the Wald 95% confidence interval for π is

$$.86 \pm 1.96(.0491) = .86 \pm .096 = (.764, .956)$$

Using the WAC confidence interval, we need to compute

$$\tilde{y} = y + .5z^2_{\alpha/2} = 43 + .5(1.96)^2 = 44.9208$$

$$\tilde{n} = n + z^2_{\alpha/2} = 50 + (1.96)^2 = 53.8416$$

and

$$\tilde{\pi} = \frac{\tilde{y}}{\tilde{n}} = \frac{44.9208}{53.8416} = .8343$$

which yields the WAC 95% confidence interval for π:

$$\left(.8343 - 1.96\sqrt{\frac{.8343(1 - .8343)}{53.8416}}, .8343 + 1.96\sqrt{\frac{.8343(1 - .8343)}{53.8416}}\right) = (.735, .934)$$

In this particular example, the traditional and WAC confidence intervals are not substantially different. However, as π approaches either 0 or 1, the difference in the two intervals can be substantial. ■

Another problem that arises in the estimation of π occurs when π is very close to 0 to 1. In these situations, the population proportion would often be estimated to be 0 or 1, respectively, unless the sample size is extremely large. These estimates are not realistic, since they would suggest that either no successes or no failures exist in the population. Rather than estimating π using the formula $\hat{\pi}$ given previously, adjustments are provided to prevent the estimates from being so extreme. One of the proposed adjustments is to use

$$\hat{\pi}_{\text{Adj.}} = \frac{\frac{3}{8}}{\left(n + \frac{3}{4}\right)} \quad \text{when } y = 0$$

and

$$\hat{\pi}_{\text{Adj.}} = \frac{\left(n + \frac{3}{8}\right)}{\left(n + \frac{3}{4}\right)} \quad \text{when } y = n$$

When computing the confidence interval for π in those situations where $y = 0$ or $y = n$, the confidence intervals using the normal approximation would not be valid. We can use the following confidence intervals, which are derived using the binomial distribution.

100(1 − α)% Confidence Interval for π When y = 0 or y = n

When $y = 0$, the confidence interval is $(0, 1 - (\alpha/2)^{1/n})$.

When $y = n$, the confidence interval is $((\alpha/2)^{1/n}, 1)$.

EXAMPLE 10.3

A new PC operating system is being developed. The designer claims the new system will be compatible with nearly all computer programs currently being run on the Microsoft Windows operating system. A sample of 50 programs is run, and all 50 programs perform without error. Estimate π, the proportion of all Microsoft Windows–compatible programs that would run without change on the new operating system. Compute a 95% confidence interval for π.

Solution If we used the standard estimator of π, we would obtain

$$\hat{\pi} = \frac{50}{50} = 1.0$$

Thus, we would conclude that 100% of all programs that are Microsoft Windows–compatible programs would run without alteration on the new operating system. Would this conclusion be valid? Probably not, since we have only investigated a tiny fraction of all Microsoft Windows–compatible programs. Thus, we will use the alternative estimators and confidence interval procedures. The point estimator would be given by

$$\hat{\pi}_{\text{Adj.}} = \frac{\left(n + \frac{3}{8}\right)}{\left(n + \frac{3}{4}\right)} = \frac{\left(50 + \frac{3}{8}\right)}{\left(50 + \frac{3}{4}\right)} = .993$$

A 95% confidence interval for π would be

$$((\alpha/2)^{1/n}, 1) = ((.05/2)^{1/50}, 1) = ((.025)^{.02}, 1) = (.929, 1.0)$$

We would now conclude that we are reasonably confident (95%) a high proportion (between 92.9% and 100%) of all programs that are Microsoft Windows–compatible would run without alteration on the new operating system. ■

Keep in mind, however, that a sample size that is sufficiently large to satisfy the rule *does not* guarantee that the interval will be informative. It only judges the adequacy of the normal approximation to the binomial—the basis for the confidence level.

Sample-size calculations for estimating π follow very closely the procedures we developed for inferences about μ. The required sample size for a $100(1 - \alpha)\%$ confidence interval for π of the form $\hat{\pi} \pm E$ (where E is specified) is found by solving the expression

$$z_{\alpha/2}\sigma_{\hat{\pi}} = E$$

for n. The result is shown here.

Sample Size Required for a 100(1 − α)% Confidence Interval for π of the Form $\hat{\pi} \pm E$

$$n = \frac{z^2_{\alpha/2}\pi(1 - \pi)}{E^2}$$

Note: Since π is not known, either substitute an educated guess or use $\pi = .5$. Use of $\pi = .5$ will generate the largest possible sample size for the specified confidence interval width, $2E$, and thus will give a conservative answer to the required sample size.

EXAMPLE 10.4

The designer of the new operating system introduced in Example 10.3 has decided to conduct a more extensive study. She wants to determine how many programs to randomly sample in order to estimate the proportion of Microsoft Windows–compatible programs that would perform adequately using the new operating system. The designer wants the estimator to be within .03 of the true proportion using a 95% confidence interval as the estimator.

Solution The designer wants the 95% confidence interval to be of the form $\hat{\pi} \pm .03$. The sample size necessary to achieve this accuracy is given by

$$n = \frac{z_{\alpha/2}^2 \pi(1 - \pi)}{E^2}$$

where the specification of 95% yields $z_{\alpha/2} = z_{.025} = 1.96$ and $E = .03$. If we did not have any prior information about π, then $\pi = .5$ must be used in the formula, yielding

$$n = \frac{(1.96)^2.5(1 - .5)}{(.03)^2} = 1{,}067.1$$

That is, 1,068 programs would need to be tested in order to be 95% confident that the estimate of π is within .03 of the actual value of π. The lower bound of the estimate of π obtained in Example 10.3 was .929. Suppose the designer is not too confident in this value but fairly certain that π is greater than .80. Using $\pi = .8$ as a lower bound, then the value of n is given by

$$n = \frac{(1.96)^2.8(1 - .8)}{(.03)^2} = 682.95$$

Thus, if the designer is fairly certain that the actual value of π is at least .80, then the required sample size can be greatly reduced, from 1,068 to 683. ■

A statistical test about a binomial parameter π is very similar to the large-sample test concerning a population mean presented in Chapter 5. These results are summarized next, with three different alternative hypotheses along with their corresponding rejection regions. Recall that only one alternative is chosen for a particular problem.

Summary of a Statistical Test for π, π_0 is Specified

H_0: **1.** $\pi \leq \pi_0$ **2.** $\pi \geq \pi_0$ **3.** $\pi = \pi_0$

H_a: **1.** $\pi > \pi_0$ **2.** $\pi < \pi_0$ **3.** $\pi \neq \pi_0$

T.S.: $z = \dfrac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}}$

R.R.: For a probability α of a Type I error

1. Reject H_0 if $z > z_\alpha$.

2. Reject H_0 if $z < -z_\alpha$.

3. Reject H_0 if $|z| > z_{\alpha/2}$.

Note: Under H_0,

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$$

Also, n must satisfy both $n\pi_0 \geq 5$ and $n(1 - \pi_0) \geq 5$.

Check assumptions and draw conclusions.

EXAMPLE 10.5

One of the largest problems on college campuses is alcohol abuse by underage students. Although all 50 states have mandated by law that no one under the age of 21 may possess or purchase alcohol, many college students report that alcohol is readily available. More problematic is that these same students report that they drink with one goal in mind—to get drunk. Universities are acutely aware of the problem of binge drinking, defined as consuming five or more drinks in a row three or more times in a 2-week period. An extensive survey of college students reported that 44% of U.S. college students engaged in binge drinking during the 2 weeks before the survey. The president of a large midwestern university stated publicly that binge drinking was not a problem on her campus of 25,000 undergraduate students. A service fraternity conducted a survey of 2,500 undergraduates attending the university and found that 1,200 of the 2,500 students had engaged in binge drinking. Is there sufficient evidence to indicate that the percentage of students engaging in binge drinking at the university is greater than the percentage found in the national survey? Use $\alpha = .05$ and also place a 95% confidence interval on the percentage of binge drinkers at the university.

Solution Let π be the proportion of undergraduates at the university that binge drink. The hypotheses of interest are

$$H_0: \ \pi \leq .44 \ \text{ versus } \ H_a: \pi > .44$$

$$\text{T.S.: } z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}}$$

R.R.: For $\alpha = .05$, reject H_0 if $z > 1.645$.

From the survey data, calculate

$$\hat{\pi} = \frac{1{,}200}{2{,}500} = .48 \quad \text{and} \quad \sigma_{\hat{\pi}} = \sqrt{\frac{(.44)(1 - .44)}{2{,}500}} = .009928$$

Also,

$$n\pi_0 = 2{,}500(.44) = 1{,}100 > 5 \ \text{ and } \ n(1 - \pi_0) = 2{,}500(1 - .44) = 1{,}400 > 5$$

Thus, the large-sample z is valid, and we obtain

$$z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}} = \frac{.48 - .44}{.009928} = 4.00 > 1.645$$

Because the observed value of z exceeds the critical value 1.645, we conclude there is significant evidence that the percentage of students that participate in binge drinking

exceeds the national percentage of 44%. The strength of the evidence is given by p-value $= Pr[z > 4.00] = .00003$. A 95% confidence interval for π is given by

$$\tilde{n} = 2{,}500 + (1.96)^2 = 2{,}503.84 \qquad \tilde{\pi} = \frac{1{,}200 + .5(1.96)^2}{2{,}503.84} = .4800$$

$$.48 \pm 1.96\sqrt{\frac{.48(1 - .48)}{2{,}503.84}} = .48 \pm .0196 = (.46, .50)$$

Thus, the percentage of binge drinkers at the university is, with 95% confidence, between 46% and 50%. ■

When either $n\pi_0 < 5$ or $n(1 - \pi_0) < 5$, the distribution of the test statistic $z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}}$ will be skewed. Thus, the normal approximation will not provide accurate values for the critical value or for the p-values. In these situations, an exact binomial test can be implemented.

In n trials of a binomial experiment, suppose we observe y successes. Our estimate of π is $\hat{\pi} = y/n$. Now suppose we want to test hypotheses comparing the binomial proportion π to a claimed value π_0. Our test statistic is Y, which has a binomial distribution with parameters n and π_0. The following display will illustrate how to obtain the p-value for various tests of hypotheses.

Summary of the Binomial Test for π

H_0:		H_a:	
1.	$\pi \leq \pi_0$	1.	$\pi > \pi_0$
2.	$\pi \geq \pi_0$	2.	$\pi < \pi_0$
3.	$\pi = \pi_0$	3.	$\pi \neq \pi_0$

T.S. Y distributed binomial (n, π_0):

1. p-value $= P(Y \geq y) = 1 - P(Y \leq y - 1) = 1 - \textbf{\textit{pbinom}}(y - 1, n, \pi_0)$
2. p-value $= P(Y \leq y) = \textbf{\textit{pbinom}}(y, n, \pi_0)$

3a. If $\hat{\pi} \geq \pi_0$, then
p-value $= 2P(Y \geq y) = 2(1 - P(Y \leq y - 1)) = 2(1 - \textbf{\textit{pbinom}}(y - 1, n, \pi_0))$

3b. If $\hat{\pi} < \pi_0$, then
p-value $= 2P(Y \leq y) = \textbf{2}\textbf{\textit{pbinom}}(y, n, \pi_0)$

R.R.: In all three cases, reject H_0 if p-value $\leq \alpha$.

- The value of $\textbf{\textit{pbinom}}(y, n, \pi_0)$ can be calculated using formulas from Chapter 4 or from a software packages such as R.

Example 10.6

The public health department in a county with a large number of oil wells was been assigned the task of evaluating whether the wastewater from the oil wells has polluted the water from private water wells near the drilling sites. In a preliminary study, a random sample of 15 oil wells was selected in the county. For each of the selected oil wells, a water well within .25 kilometers of the oil well is examined. In 4 of the 15 wells, the level of endrocrine-disrupting chemicals was above the level that can cause interferences with the body's normal hormonal function. These chemicals

are known to occur naturally in approximately 20% of water wells. Is there significant evidence that more than 20% of the water wells near an oil well are contaminated with endrocrine-disrupting chemicals?

Solution Let π be the proportion of water wells near oil wells that are contaminated with endrocrine-disrupting chemicals. The hypotheses of interest are

$$H_0: \pi \leq .20 \quad \text{versus} \quad H_a: \pi > .20$$

Note: $n\pi_0 = (15)(.2) = 3 < 5$. Thus, the normal-based z test would not be appropriate to test the hypotheses.

From the sampled wells, $Y = 4$ of the 15 wells were contaminated. Under H_0, Y has a binomial distribution with $n = 15$ and $\pi = \pi_0 = .2$. Using the formula for computing binomial probabilities from Chapter 4, the p-value is computed to be

$$\begin{aligned} p\text{-value} &= P(Y \geq 4) = 1 - P(Y \leq 3) \\ &= 1 - [P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3] \\ &= 1 - [.0352 + .1319 + .2309 + .2501] = .3519 \end{aligned}$$

Alternatively, using the R binomial function, p-value $= 1 -$ ***pbinom***$(3, 15, .2) =$ .3518.

We can thus conclude that based on the small sample size, there is not significant evidence that more than 20% of the water wells near oil wells in this county are contaminated with endrocrine-disrupting chemicals even though $\hat{\pi} = \frac{4}{15} = .27 > .2$.

A 95% confidence interval for π would be obtained as follows:

$$\tilde{y} = 4 + .5(1.96)^2 = 5.9208; \quad \tilde{n} = 15 + (1.96)^2 = 18.8416; \quad \tilde{\pi} = \frac{5.9208}{18.8416} = .3142$$

$$\left(.3142 - 1.96\sqrt{\frac{.3142(1 - .3412)}{18.8416}}, .3142 + 1.96\sqrt{\frac{.3142(1 - .3412)}{18.8416}}\right) = (.105, .524)$$

The 95% confidence interval for the proportion of contaminated wells is very wide, which reflects the small sample size in the study. Even though there was not significant evidence in the observed data that more than 20% of water wells were contaminated, the 95% confidence interval induced the county to plan a much larger study. ■

10.3 Inferences About the Difference Between Two Population Proportions, $\pi_1 - \pi_2$

Many practical problems involve the comparison of two binomial parameters. Social scientists may wish to compare the proportions of women who take advan-tage of prenatal health services in two communities representing different socio-economic backgrounds. A director of marketing may wish to compare the public awareness of a new product recently launched and that of a competitor's product.

For comparisons of this type, we assume that independent random samples are drawn from two binomial populations with unknown parameters designated by π_1

and π_2. If y_1 successes are observed for the random sample of size n_1 from population 1 and y_2 successes are observed for the random sample of size n_2 from population 2, then the point estimates of π_1 and π_2 are the observed sample proportions $\hat{\pi}_1$ and $\hat{\pi}_2$, respectively.

$$\hat{\pi}_1 = \frac{y_1}{n_1} \text{ and } \hat{\pi}_2 = \frac{y_2}{n_2}$$

This notation is summarized next.

Notation for Comparing Two Binomial Proportions

	Population 1	Population 2
Population proportion	π_1	π_2
Sample size	n_1	n_2
Number of successes	y_1	y_2
Sample proportion	$\hat{\pi}_1 = \frac{y_1}{n_1}$	$\hat{\pi}_2 = \frac{y_2}{n_2}$

Inferences about two binomial proportions are usually phrased in terms of their difference, $\pi_1 - \pi_2$, and we use the difference in sample proportions, $\hat{\pi}_1 - \hat{\pi}_2$, as part of a confidence interval or statistical test. The sampling distribution for $\hat{\pi}_1 - \hat{\pi}_2$ can be approximated by a normal distribution with mean and standard error given by

$$\mu_{\hat{\pi}_1 - \hat{\pi}_2} = \pi_1 - \pi_2$$

and

$$\sigma_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}$$

rule for sample sizes

This approximation is appropriate if we apply the same requirements to both binomial populations that we applied in recommending a normal approximation to a binomial (see Chapter 4). Thus, the normal approximation to the distribution of $\hat{\pi}_1 - \hat{\pi}_2$ is appropriate if both $n_i\pi_i$ and $n_i(1 - \pi_i)$ are 5 or more for $i = 1, 2$. Since π_1 and π_2 are not known, the validity of the approximation is determined by examining $n_i\hat{\pi}_i$ and $n_i(1 - \hat{\pi}_i)$ for $i = 1, 2$.

Confidence intervals and statistical tests about $\pi_1 - \pi_2$ are straightforward and follow the format we used for comparisons using $\mu_1 - \mu_2$. Interval estimation is summarized here; it takes the usual form, point estimate $\pm$ z (standard error).

100(1 − α)% Confidence Interval for $\pi_1 - \pi_2$

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\sigma_{\hat{\pi}_1 - \hat{\pi}_2}$$

where

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}$$

EXAMPLE 10.7

A company test-markets a new product in the Grand Rapids, Michigan, and Wichita, Kansas, metropolitan areas. The company's advertising in the Grand Rapids area is based almost entirely on television commercials. In Wichita, the company spends a roughly equal dollar amount on a balanced mix of television, radio, newspaper, and magazine ads. Two months after the ad campaign begins, the company conducts surveys to determine consumer awareness of the product.

TABLE 10.1
Survey data for example.

	Grand Rapids	Wichita
Number interviewed	608	527
Number aware	392	413

Calculate a 95% confidence interval for the regional difference in the proportions of all consumers who are aware of the product (as shown in Table 10.1).

Solution The sample awareness proportion is higher in Wichita, so let's make Wichita region 1.

$$\hat{\pi}_1 = 413/527 = .784 \quad \hat{\pi}_2 = 392/608 = .645$$

The estimated standard error is

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{(.784)(.216)}{527} + \frac{(.645)(.355)}{608}} = .0264$$

Therefore, the 95% confidence interval is

$$(.784 - .645) \pm 1.96(.0264) = (.087, .191)$$

which indicates that somewhere between 8.7% and 19.1% more Wichita consumers than Grand Rapids consumers are aware of the product. ■

This confidence interval method is based on the normal approximation to the binomial distribution. In Chapter 4, we indicated as a general rule that $n\hat{\pi}$ and $n(1 - \hat{\pi})$ should both be at least 5 to use this normal approximation. For this confidence interval to be used, the sample size rule should hold for each sample.

The reason for confidence intervals that seem very wide and unhelpful is that each measurement conveys very little information. In effect, each measurement conveys only one "bit": a 1 for a success or a 0 for a failure. For example, surveys of the compensation of chief executive officers of companies often give a manager's age in years. If we replaced the actual age by a category such as "over 55 years old" versus "under 55," we definitely would have far less information. When there is little information per item, we need a large number of items to get an adequate total amount of information. Wherever possible, it is better to have a genuinely numerical measure of a result rather than mere categories. When numerical measurement isn't possible, relatively large sample sizes will be needed.

Hypothesis testing about the difference between two population proportions is based on the z statistic from a normal approximation. The typical null hypothesis is that there is no difference between the population proportions, though any specified value for $\pi_1 - \pi_2$ may be hypothesized. The procedure is very much like a t test of the difference of means and is summarized next.

Statistical Test for the Difference between Two Population Proportions

H_0: **1.** $\pi_1 - \pi_2 \leq 0$ $\quad$ H_a: **1.** $\pi_1 - \pi_2 > 0$
2. $\pi_1 - \pi_2 \geq 0$ $\quad$ **2.** $\pi_1 - \pi_2 < 0$
3. $\pi_1 - \pi_2 = 0$ $\quad$ **3.** $\pi_1 - \pi_2 \neq 0$

T.S.: $$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2)}{\sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}}$$

R.R.: **1.** $z > z_\alpha$.
2. $z < -z_\alpha$.
3. $|z| > z_{\alpha/2}$.

Check assumptions and draw conclusions.

Note: This test should be used only if $n_1\hat{\pi}_1$, $n_1(1 - \hat{\pi}_1)$, $n_2\hat{\pi}_2$, and $n_2(1 - \hat{\pi}_2)$ are all at least 5.

EXAMPLE 10.8

An educational researcher designs a study to compare the effectiveness of teaching English to non-English-speaking people by a computer software program and by the traditional classroom system. The researcher randomly assigns 125 students from a class of 300 to instruction using the computer. The remaining 175 students are instructed using the traditional method. At the end of a 6-month instructional period, all 300 students are given an examination with the results reported in Table 10.2.

TABLE 10.2
Exam data for example

Exam Result	Computer Instruction	Traditional Instruction
Pass	94	113
Fail	31	62
Total	125	175

Does instruction using the computer software program appear to increase the proportion of students passing the examination in comparison to the pass rate using the traditional method of instruction? Use $\alpha = .05$.

Solution Denote the proportion of all students passing the examination using the computer method of instruction and the traditional method of instruction by π_1 and π_2, respectively. We will test the hypotheses

$$H_0: \ \pi_1 - \pi_2 \leq 0$$
$$H_a: \ \pi_1 - \pi_2 > 0$$

We will reject H_0 if the test statistic z is greater than $z_{.05} = 1.645$. From the data, we compute the estimates

$$\hat{\pi}_1 = \frac{94}{125} = .752 \text{ and } \hat{\pi}_2 = \frac{113}{175} = .646$$

From these, we compute the test statistic to be

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} = \frac{.752 - .646}{\sqrt{\frac{.752(1-.752)}{125} + \frac{.646(1-.646)}{175}}} = 2.00$$

Since $z = 2.00$ is greater than 1.645, we reject H_0 and conclude that the observations support the hypothesis that the computer instruction results in a higher pass rate than the traditional approach. The p-value of the observed data is given by p-value $= P(z \geq 2.00) = .0228$, using the standard normal tables. A 95% confidence interval on the effect size $\pi_1 - \pi_2$ is given by

$$.752 - .646 \pm 1.96\sqrt{\frac{.752(1-.752)}{125} + \frac{.646(1-.646)}{175}} = .106 \pm .104 = (.002, .210)$$

We are 95% confident that the proportion passing the examination is between .2% and 21% higher for students using computer instruction than those using the traditional approach. For our conclusions to have a degree of validity, we need to check whether the sample sizes were large enough. Now, $n_1\hat{\pi}_1 = 94$, $n_1(1-\hat{\pi}_1) = 31$, $n_2\hat{\pi}_2 = 113$, and $n_2(1-\hat{\pi}_2) = 62$; thus, all four quantities are greater than 5. Hence, the large-sample criterion would appear to be satisfied. ■

Fisher Exact test

When at least one of the conditions—$n_1\hat{\pi}_1 \geq 5$, $n_1(1-\hat{\pi}_1) \geq 5$, $n_2\hat{\pi}_2 \geq 5$, or $n_2(1-\hat{\pi}_2) \geq 5$—for using the large-sample approximation to the distribution of the test statistic for comparing two proportions is invalid, the **Fisher Exact test** should be used.

The hypotheses to be tested are H_0: $\pi_1 \leq \pi_2$ versus H_a: $\pi_1 > \pi_2$, where π_is are the probabilities of "success" for populations $i = 1, 2$. In developing a small-sample test of hypotheses, we need to develop the exact probability distribution for the cell counts in all 2×2 tables having the same row and column totals as the 2×2 table from the observed data (Table 10.3).

TABLE 10.3
Cell counts in 2×2 table

	Outcome		
Population	**Success**	**Failure**	**Total**
1	x	$n_1 - x$	n_1
2	y	$n_2 - y$	n_2
Total	m	$n - m$	n

For tables having the same row and column totals—n_1, n_2, m, and $n - m$—the value of x determines the counts for the remaining three cells because $y = m - x$.

When $\pi_1 = \pi_2$, the probability of observing a particular value for x—that is, the probability of a particular table being observed—is given by

$$P(x = k) = \frac{\binom{n_1}{k}\binom{n_2}{m-k}}{\binom{n}{m}}$$

where

$$\binom{n_1}{k} = \frac{n_1(n_1-1)(n_1-2)\cdots(n_1-k+1)}{k(k-1)(k-2)\cdots 1}$$

To test the difference in the two population proportions, the p-value of the test is the sum of these probabilities for outcomes at least as supportive the alternative hypothesis as the observed table. For H_a: $\pi_1 > \pi_2$, we need to determine which other possible 2×2 tables would provide stronger support of H_a than the observed table. Given the marginal totals—n_1, n_2, m, and $n - m$—tables having larger x values will have larger values for $\hat{\pi}_1$ and hence provide stronger evidence in favor of $\pi_1 > \pi_2$.

The possible values of x are $0, 1, \ldots, \min(n_1, m)$ and hence

$$p\text{-value} = P[x \geq k] = \sum_{j=k}^{\min(n_1,m)} \frac{\binom{n_1}{j}\binom{n_2}{m-j}}{\binom{n}{m}}$$

For the two-sided alternative, H_a: $\pi_1 \neq \pi_2$, the p-value is defined as the sum of the probabilities of tables no more likely than the observed table. Thus, the p-value is the sum of the probabilities of all values of $x = j$ for which $P(j) \leq P(k)$, where k is the observed value of x. We will illustrate these calculations with the following example.

EXAMPLE 10.9

A clinical trial is conducted to compare two drug therapies for leukemia: P and PV. Twenty-one patients were assigned to drug P and 42 patients to drug PV. Table 10.4 summaries the success of the two drugs:

TABLE 10.4
Outcomes of drug therapies

	Outcome		
Drug	**Success**	**Failure**	**Total**
PV	38	4	42
P	14	7	21
Total	52	11	63

Is there significant evidence that the proportion of patients obtaining a successful outcome is higher for drug PV than for drug P?

Solution First, we check the conditions for using the large-sample test:

$$n_1\hat{\pi}_1 = 38 \geq 5, \quad n_1(1 - \pi_1) = 4 < 5, \quad n_2\pi_2 = 14 \geq 5, \quad n_2(1 - \pi_2) = 7 > 5$$

Because one of the four conditions is violated, the large-sample test should not be applied.

The Fisher Exact test will be applied to this data set. First, we will compute the p-value for testing the hypotheses H_0: $\pi_P \geq \pi_{PV}$ versus H_a: $\pi_P < \pi_{PV}$. After obtaining the p-value, we will compare its value to α.

The probability of the observed table is

$$P(x = 38) = \frac{\binom{42}{38}\binom{21}{14}}{\binom{63}{52}} = .0211$$

Thus, the one-sided p-value is the sum of the probabilities for all tables having 38 or more successes:

$$p\text{-value} = P[x \geq 38] = P(x = 38) + P(x = 39) + P(x = 40) + P(x = 41) + P(x = 42)$$

$$= \frac{\binom{42}{38}\binom{21}{14}}{\binom{63}{52}} + \frac{\binom{42}{39}\binom{21}{13}}{\binom{63}{52}} + \frac{\binom{42}{40}\binom{21}{12}}{\binom{63}{52}} + \frac{\binom{42}{41}\binom{21}{11}}{\binom{63}{52}} + \frac{\binom{42}{42}\binom{21}{10}}{\binom{63}{52}}$$

$$= .02114 + .00379 + .00041 + .00002 + .00000 = .02536$$

For all values of $\alpha \leq .025$ then, the p-value $= .02536 > \alpha$, so we conclude that there is not significant evidence that the proportion of patients obtaining a successful outcome is higher for drug PV than for drug P.

If the large-sample z test would have been applied to this data set, a value of $z = 2.119$ would have been obtained with p-value $= .017$. Thus, the z test and Fisher Exact test would have yielded contradictory conclusions for values of α in the range $.017 < \alpha < .025$.

Many software packages have the Fisher Exact test as an option for testing hypotheses about two proportions. ■

The z test and the Fisher Exact test for comparing two proportions require that the two samples be independent. The McNemar test was developed for those studies in which proportions are dependent. Thus, it allows us to compare the values of two proportions that are dependent.

McNemar Test for Matched Pairs

In some situations, the information in a 2×2 contingency table is collected from experimental units for which two related responses are obtained. There are no longer n independent responses categorized into the four cells but rather a pair of responses from related units. For example, responses from the same individual at two different times (before and after an intervention) or from two individuals who are physically related (husband–wife or twins) or from body parts of the same experimental unit (right hand–left hand or right eye–left eye).

The data from a study involving matched pairs has the same form as the 2×2 tables we discussed previously except now the response is recorded in such a manner that the pairing is identified. Table 10.5 is a typical summary of the data for this type of study.

The interpretation of the data in the table is as follows: n_{11} is the number of pairs with Yes for both responses, n_{21} is the number of pairs with Yes for response 1 but No for response 2, n_{12} is the number of pairs with No for response 1 but Yes for response 2, n_{22} is the number of pairs with No for both responses.

The population of responses for all such pairs has proportions given in Table 10.6.

TABLE 10.5
Sample counts

	Response 1		
Response 2	**Yes**	**No**	**Total**
Yes	n_{11}	n_{12}	$n_{1.}$
No	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

TABLE 10.6
Population proportions

	Response 1		
Response 2	**Yes**	**No**	**Total**
Yes	π_{11}	π_{12}	$\pi_{1.}$
No	π_{21}	π_{22}	$\pi_{2.}$
Total	$\pi_{.1}$	$\pi_{.2}$	1

The research question in this situation is whether the proportion of pairs responding Yes for response 1 is the same as the proportion of pairs responding Yes for response 2. The independent-samples z test and Fisher Exact test are not valid test statistics because the cell counts may be correlated due to pairing of the two responses. We want to test the hypotheses

$$H_0\colon \pi_{1.} = \pi_{.1} \text{ versus } H_a\colon \pi_{1.} \neq \pi_{.1}$$

or the corresponding one-sided hypotheses

$$H_0\colon \pi_{1.} \geq \pi_{.1} \text{ versus } H_a\colon \pi_{1.} < \pi_{.1}$$

First, note that

$$\pi_{1.} - \pi_{.1} = (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21}$$

Therefore, a test of the marginal homogeneity for the matched pairs $H_0\colon \pi_{1.} = \pi_{.1}$ is equivalent to a test of $H_0\colon \pi_{12} = \pi_{21}$. That is, are the proportions of switches from Yes to No and from No to Yes equal?

When H_0 is true, the expected values for the counts n_{12} and n_{21} should be equal. Let $m = n_{12} + n_{21}$ be the total count in the off-diagonal cells in Table 10.5. Under H_0, the allocation of the m observations to the (1,2) and (2,1) cells is a binomial experiment with probability .5 for both of the cells. McNemar (1947) used the methodology of testing hypotheses about binomial proportions to derive the following test statistic.

Summary of the McNemar Test for Comparing Two Dependent Proportions

H_0:		H_a:
1.	$\pi_{1.} \leq \pi_{.1}$	$\pi_{1.} > \pi_{.1}$
2.	$\pi_{1.} \geq \pi_{.1}$	$\pi_{1.} < \pi_{.1}$
3.	$\pi_{1.} = \pi_{.1}$	$\pi_{1.} \neq \pi_{.1}$

Note that the above tests are equivalent to comparing π_{12} to $\frac{1}{2}$. This lead McNemar to propose the following test statistics for large-sample values of m:

Case 1: Large-sample z test

When $m = n_{12} + n_{21} \geq 20$, the following z-test can be used.

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

Let z_c be the value of z computed from the data.

1. Reject H_o if $z_c \geq z_\alpha$ with p-value $= P(z \geq z_c)$.
2. Reject H_o if $z_c \leq -z_\alpha$ with p-value $= P(z \leq z_c)$.
3. Reject H_o if $|z_c| \geq z_{\alpha/2}$ with p-value $= 2P(z \geq |z_c|)$.

Case 2: Exact binomial test

When $m = n_{12} + n_{21} < 20$, the following binomial test can be used.
T.S. Y distributed binomial$(m, .5)$

***p*-values:**

1. p-value $= P(Y \geq n_{12}) = 1 - P(Y \leq n_{12} - 1) = 1 - \textbf{\textit{pbinom}}(n_{12} - 1, m, .5)$
2. p-value $= P(Y \leq n_{12}) = \textbf{\textit{pbinom}}(n_{12}, m, .5)$

3a. If $n_{12} < (n_{12} + n_{21})/2$, then

p-value $= 2P(Y \le n_{12}) = 2$***pbinom***$(n_{12}, m, .5)$

3b. If $n_{12} > (n_{12} + n_{21})/2$, then

p-value $= 2P(Y \ge n_{12}) = 2(1 - P(Y \le n_{12} - 1)) = 2(1 -$ ***pbinom***$(n_{12} - 1, m, .5))$

3c. If $n_{12} = (n_{12} + n_{21})/2$, then

p-value $= 1$

R.R.: In all cases, reject H_o if p-value $\le \alpha$.

- The value of $pbinom(y, m, .5)$ can be calculated using formulas from Chapter 4 or from a software package such as R.

The derivation of McNemar's z test follows from the one-sample z test that was discussed earlier in this chapter. The test of H_0: $\pi_{1.} = \pi_{.1}$ versus H_a: $\pi_{1.} \ne \pi_{.1}$ is equivalent to testing H_0: $\pi_{12} = .5$ versus H_a: $\pi_{12} \ne .5$. The following equivalence demonstrates the relationship between the one-sample z test and McNemar's test statistics. The derivation uses $m = n_{12} + n_{21}$.

$$Z = \frac{\hat{p}_{12} - .5}{\sqrt{\dfrac{(.5)(1 - .5)}{m}}} = \frac{\dfrac{n_{12}}{m} - .5}{\sqrt{\dfrac{(.5)(1 - .5)}{m}}} = \frac{n_{12} - .5m}{\sqrt{(.5)(1 - .5)m}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

Taking into account the correlation between $\hat{\pi}_{1.}$ and $\hat{\pi}_{.1}$, an approximate large-sample $100(1 - \alpha)\%$ confidence interval on $\pi_{1.} - \pi_{.1}$ is

$$(\hat{\pi}_{1.} - \hat{\pi}_{.1}) \pm Z_{\alpha/2}\sqrt{\frac{\hat{\pi}_{1.}(1 - \hat{\pi}_{1.}) + \hat{\pi}_{.1}(1 - \hat{\pi}_{.1}) - 2(\hat{\pi}_{11}\hat{\pi}_{22} + \hat{\pi}_{12}\hat{\pi}_{21})}{n}}$$

which simplifies to

$$(\hat{\pi}_{1.} - \hat{\pi}_{.1}) \pm Z_{\alpha/2}\frac{1}{n}\sqrt{(n_{12} + n_{21}) - \frac{1}{n}(n_{12} - n_{21})^2}$$

Example 10.10

A case-control study was conducted in which the researchers were interested in determining if there was a relationship between diabetes and chronic circulatory problems. The 180 patients having chronic circulatory problems were matched by age, gender, occupation, and ethnicity with 180 patients without chronic circulatory problems. First, 180 pairs of subjects were then asked whether they had been diagnosed as having diabetes. The data are given in Table 10.7.

TABLE 10.7

	With Circulatory Problems		
Without Circulatory Problems	**Diabetes**	**No Diabetes**	**Total**
Diabetes	79	21	100
No Diabetes	39	41	80
Total	118	62	180

The 180 pairs of subjects in the study consist of four groups:

Group	With Circulatory Problems	Without Circulatory Problems	Count
1	Diabetes - Y	Diabetes - Y	79
2	Diabetes - Y	Diabetes - N	39
3	Diabetes - N	Diabetes - Y	21
4	Diabetes - N	Diabetes - N	41

Is the proportion of Without Circulatory Problems patients having diabetes less than the proportion of With Circulatory Problems patients having diabetes?

Solution We want to test the research hypothesis that the proportion of Without Circulatory Problems patients having diabetes is less than the proportion of With Circulatory Problems patients having diabetes. That is, test the research hypothesis H_a: $\pi_{1.} < \pi_{.1}$, or, equivalently, test H_a: $\pi_{12} < \pi_{21}$.

From the data we have $\hat{\pi}_{1.} = 100/180 = .556$ and $\hat{\pi}_{.1} = 118/180 = .656$. Therefore, $\hat{\pi}_{1.} - \hat{\pi}_{.1} = .556 - .656 = -.1$. The proportion of diabetic patients without circulatory problems is 10% less than the proportion of diabetic patients with circulatory problems. We want to next confirm this observation by applying the McNemar test to the data. Because $n_{12} + n_{21} = 21 + 39 = 60 \geq 20$, the large-sample z test is appropriate.

Our decision would be to reject H_0 if $z \leq z_{.05} = -1.645$. From the data, we have

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{21 - 39}{\sqrt{21 + 39}} = -2.324 < -1.645$$

The p-value of the test is $p\text{-value} = P(z < -2.324) = .0101$.

Our conclusion is to reject H_0 and state there is significant evidence (p-value = .0101) that the proportion of diabetes patients without circulatory problems is less than the proportion of diabetes patients with circulatory problems.

An approximate 95% confidence interval on $\pi_{1.} - \pi_{.1}$ is computed as follows:

$$(\hat{\pi}_{1.} - \hat{\pi}_{.1}) \pm Z_{\alpha/2}\frac{1}{n}\sqrt{(n_{12} + n_{21}) - \frac{1}{n}(n_{12} - n_{21})^2}$$

$$\left(\frac{100}{180} - \frac{118}{180}\right) \pm 1.96\frac{1}{180}\sqrt{(21 + 39) - \frac{1}{180}(21 - 39)^2}$$

That is, $-.10 \pm .083 = (-.183, -.017)$.

Although the sample sizes were such that the large-sample test could be applied, we will illustrate the binomial version of the McNemar test next.

The p-value $= P[Y \leq 21]$, where $m = 21 + 39 = 60$ and Y is Bin(60, .5). Thus, p-value $= P[Y \leq 21] = 0.0137 < .05$, which implies that we should reject H_0. Hence, our conclusion is the same as the conclusion from the z test with the exception being that the p-value from the binomial version of the McNemar test is slightly larger than the value from the z test. ■

10.4 Inferences About Several Proportions: Chi-Square Goodness-of-Fit Test

We can extend the binomial sampling scheme of Chapter 4 to situations in which each trial results in one of k possible outcomes ($k > 2$). For example, a random sample of registered voters is classified according to political party (Republican, Democrat, Socialist, Green, Independent, etc.) or patients in a clinical trial are evaluated with respect to the degree of improvement in their medical condition (substantially improved, improved, no change, worse). This type of experiment or study is called a multinomial experiment and has the characteristics listed here.

The Multinomial Experiment

1. The experiment consists of n identical trials.
2. Each trial results in one of k outcomes.
3. The probability that a single trial will result in outcome i is π_i for $i = 1, 2, \ldots, k$, and remains constant from trial to trial. (*Note:* $\sum \pi_i = 1$.)
4. The trials are independent.
5. We are interested in n_i, the number of trials resulting in outcome i. (*Note:* $\sum n_i = n$.)

multinomial distribution

The probability distribution for the number of observations resulting in each of the k outcomes, called the **multinomial distribution,** is given by the formula

$$P(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_k^{n_k}$$

Recall from Chapter 4, where we discussed the binomial probability distribution, that

$$n! = n(n-1)\cdots 1$$

and

$$0! = 1$$

We can use the formula for the multinomial distribution to compute the probability of particular events.

EXAMPLE 10.11

Previous experience with the breeding of a particular herd of cattle suggests that the probability of obtaining one healthy calf from a mating is .83. Similarly, the probabilities of obtaining zero or two healthy calves are, respectively, .15 and .02. A farmer breeds three dams from the herd; find the probability of obtaining exactly three healthy calves.

Solution Assuming the three dams are chosen at random, this experiment can be viewed as a multinomial experiment with $n = 3$ trials and $k = 3$ outcomes. These outcomes are listed in Table 10.8 with the corresponding probabilities.

TABLE 10.8 Probabilities of progeny occurrences

Outcome	Number of Progeny	Probability, π_i
1	0	.15
2	1	.83
3	2	.02

Note that outcomes 1, 2, and 3 refer to the events that a dam produces zero, one, or two healthy calves, respectively. Similarly, n_1, n_2, and n_3 refer to the number of dams producing zero, one, or two healthy progeny, respectively. To obtain exactly three healthy progeny, we must observe one of the following possible events.

$$\text{A: } \begin{cases} \text{1 dam gives birth to no healthy progeny: } n_1 = 1 \\ \text{1 dam gives birth to 1 healthy progeny: } n_2 = 1 \\ \text{1 dam gives birth to 2 healthy progeny: } n_3 = 1 \end{cases}$$

$$\text{B: 3 dams give birth to 1 healthy progeny: } \begin{cases} n_1 = 0 \\ n_2 = 3 \\ n_3 = 0 \end{cases}$$

For event A with $n = 3$ and $k = 3$,

$$P(n_1 = 1, n_2 = 1, n_3 = 1) = \frac{3!}{1!1!1!}(.15)^1(.83)^1(.02)^1 \cong .015$$

Similarly, for event B,

$$P(n_1 = 0, n_2 = 3, n_3 = 0) = \frac{3!}{0!3!0!}(.15)^0(.83)^3(.02)^0 = (.83)^3 \cong .572$$

Thus, the probability of obtaining exactly three healthy progeny from three dams is the sum of the probabilities for events A and B; namely, .015 + .572 = .587. ■

Our primary interest in the multinomial distribution is as a probability model underlying statistical tests about the probabilities $\pi_1, \pi_2, \ldots, \pi_k$. We will hypothesize specific values for the πs and then determine whether the sample data agree with the hypothesized values. One way to test such a hypothesis is to examine the observed number of trials resulting in each outcome and to compare this to the number we would *expect* to result in each outcome. For instance, in our previous example, we gave the probabilities associated with zero, one, and two progeny as .15, .83, and .02. In a sample of 100 mated dams, we would **expect to observe** 15 dams that produce no healthy progeny. Similarly, we would expect to observe 83 dams that produce one healthy calf and 2 dams that produce two healthy calves.

expected number of outcomes

DEFINITION 10.1

In a multinomial experiment in which each trial can result in one of k outcomes, the **expected number of outcomes** of type i in n trials is $n\pi_i$, where π_i is the probability that a single trial results in outcome i.

In 1900, Karl Pearson proposed the following test statistic to test the specified probabilities:

$$\chi^2 = \sum_i \left[\frac{(n_i - E_i)^2}{E_i} \right]$$

where n_i represents the number of trials resulting in outcome i and E_i represents the number of trials we would expect to result in outcome i when the hypothesized probabilities represent the actual probabilities assigned to each outcome. Frequently, we will refer to the probabilities $\pi_1, \pi_2, \ldots, \pi_k$ as **cell probabilities,** one cell corresponding to each of the k outcomes. The observed numbers $n_1, n_2, \ldots, n_k$ corresponding to the k outcomes will be called **observed cell counts,** and the expected numbers $E_1, E_2, \ldots, E_k$ will be referred to as **expected cell counts.**

cell probabilities

observed cell counts
expected cell counts

Suppose that we hypothesize values for the cell probabilities $\pi_1, \pi_2, \ldots, \pi_k$. We can then calculate the expected cell counts by using Definition 10.1 to examine how well the observed data fit, or agree, with what we would expect to observe. Certainly, if the hypothesized π-values are correct, the observed cell counts, n_i, should not deviate greatly from the expected cell counts, E_i, and the computed value of χ^2 should be small. Similarly, when one or more of the hypothesized cell probabilities are incorrect, the observed and expected cell counts will differ substantially, making χ^2 large.

chi-square distribution

The distribution of the quantity χ^2 can be approximated by a **chi-square distribution** provided that the expected cell counts, E_i, are fairly large.

The chi-square goodness-of-fit test based on k specified cell probabilities will have $k - 1$ degrees of freedom. We will explain why we have $k - 1$ degrees of freedom at the end of this section. Upper-tail values of the test statistic

$$\chi^2 = \sum_i \left[\frac{(n_i - E_i)^2}{E_i} \right]$$

can be found in Table 7 in the Appendix.

We can now summarize the chi-square goodness-of-fit test concerning k specified cell probabilities.

Chi-Square Goodness-of-Fit Test

H_0: $\pi_i = \pi_{i0}$ for categories $i = 1,\ldots,k$, π_{i0} are specified probabilities or proportions.

H_a: At least one of the cell probabilities differs from the hypothesized value.

T.S.: $\chi^2 = \sum \left[\frac{(n_i - E_i)^2}{E_i} \right]$

where n_i is the observed number in category i and $E_i = n\pi_{i0}$ is the expected number under H_0.

R.R.: Reject H_0 if χ^2 exceeds the tabulated critical value for the specified α and df $= k - 1$.

Check assumptions and draw conclusions.

The approximation of the sampling distribution of the chi-square goodness-of-fit test statistic by a chi-square distribution improves as the sample size n becomes larger. The accuracy of the approximation depends on both the sample size n and the number of cells k. Cochran (1954) indicates that the approximation should be adequate if no E_i is less than 1 and no more than 20% of the E_is are less than 5. The values of n/k that provide adequate approximations for the chi-square goodness-of-fit test statistic tends to decrease as k increases. Agresti (2002) discusses situations in which the chi-squared approximation tends to be poor for studies having small observed cell counts even if the expected cell counts are moderately large. Agresti concludes that it is hopeless to determine a single rule concerning the appropriate sample size to cover all cases. However, we recommend applying Cochran's guidelines for determining whether the chi-square goodness-of-fit test statistic can be adequately approximated with a chi-square distribution. When some of the E_is are too small, there are several alternatives. Researchers combine levels of the categorical variable to increase the observed cell counts. However, combining categories should not be done unless there is a natural way

to redefine the levels of the categorical variable that does not change the nature of the hypothesis to be tested. When it is not possible to obtain observed cell counts large enough to permit the chi-squared approximation, Agresti (2002) discusses *exact* methods to test the hypotheses. Many software pakages include these exact tests as an option.

EXAMPLE 10.12

A laboratory is comparing a test drug to a standard drug preparation that is useful in the maintenance of patients suffering from high blood pressure. Over many clinical trials at many different locations, the standard therapy was administered to patients with comparable hypertension (as measured by the New York Heart Association (NYHA) Classification). The lab then classified the responses to therapy for this large patient group into one of four response categories. Table 10.9 lists the categories and percentages of patients treated using the standard preparation who have been classified in each category.

TABLE 10.9
Results of clinical trials using the standard preparation

Category	Percentage
Marked decrease in blood pressure	50
Moderate decrease in blood pressure	25
Slight decrease in blood pressure	10
Stationary or slight increase in blood pressure	15

The lab then conducted a clinical trial with a random sample of 200 patients with high blood pressure. All patients were required to be listed according to the same hypertensive categories of the NYHA Classification as those studied under the standard preparation. Use the sample data in Table 10.10 to test the hypothesis that the cell probabilities associated with the test preparation are identical to those for the standard. Use $\alpha = .05$.

TABLE 10.10
Sample data for example

Category	Observed Cell Counts
1	120
2	60
3	10
4	10

Solution This experiment possesses the characteristics of a multinomial experiment with $n = 200$ and $k = 4$ outcomes.

Outcome 1: A person's blood pressure will decrease markedly after treatment with the test drug.
Outcome 2: A person's blood pressure will decrease moderately after treatment with the test drug.
Outcome 3: A person's blood pressure will decrease slightly after treatment with the test drug.
Outcome 4: A person's blood pressure will remain stationary or increase slightly after treatment with the test drug.

The null and alternative hypotheses are then

$$H_0: \pi_1 = .50, \pi_2 = .25, \pi_3 = .10, \pi_4 = .15$$

and

H_a: At least one of the cell probabilities is different from the hypothesized value.

Before computing the test statistic, we must determine the expected cell numbers. These data are given in Table 10.11.

TABLE 10.11
Observed and expected cell numbers for example

Category	Observed Cell Number, n_i	Expected Cell Number, E_i
1	120	200(.50) = 100
2	60	200(.25) = 50
3	10	200(.10) = 20
4	10	200(.15) = 30

Because all the expected cell numbers are relatively large, we may calculate the chi-square statistic and compare it to a tabulated value of the chi-square distribution.

$$\chi^2 = \sum_i \left[\frac{(n_i - E_i)^2}{E_i}\right]$$
$$= \frac{(120 - 100)^2}{100} + \frac{(60 - 50)^2}{50} + \frac{(10 - 20)^2}{20} + \frac{(10 - 30)^2}{30}$$
$$= 4 + 2 + 5 + 13.33 = 24.33$$

For the probability of a Type I error set at $\alpha = .05$, we look up the value of the chi-square statistic for $a = .05$ and df $= k - 1 = 3$. The critical value from Table 7 in the Appendix is 7.815.

R.R.: Reject H_0 if $\chi^2 > 7.815$.

Conclusion: The computed value of χ^2 is greater than 7.815, so we reject the null hypothesis and conclude there is significant evidence that at least one of the cell probabilities differs from that specified under H_0. Practically, it appears that a much higher proportion of patients treated with the test preparation falls into the moderate and marked improvement categories. The p-value for this test is $p < .001$. (See Table 7 in the Appendix, or use R function $1 -$ ***pchisq***$(24.33, 3) = .00002$.) ■

Goodness-of-Fit of a Probability Model

In situations in which a researcher has count data—for example, number of a particular insect on randomly selected plants or number of times a particular event occurs in a fixed period of time—the researcher may want to determine if a particular probability model adequately fits the data. Does a binomial or Poisson model provide a reasonable model for the observed data? The measure of how well the data fit the model is the chi-square goodness-of-fit statistic:

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(n_i - E_i)^2}{E_i}\right]$$

In the chi-square goodness-of-fit statistic, the quantity n_i denotes the number of observations in cell i, and E_i is the expected number in cell i assuming the proposed model is correct. We will illustrate the procedures used to check the adequacy of a proposed probability model using the Poisson distribution.

There are two types of hypotheses. The first type of hypothesis has a completely specified model for the data. The hypothesis is that the data arise from a Poisson distribution with $\mu = \mu_0$, where μ_0 is specified by the researcher. The hypotheses being tested are

H_0: Data arise from a Poisson model with $\mu = \mu_0$.

H_a: Data do not arise from a Poisson model.

In this situation, the E_is are computed from a Poisson model with $\mu = \mu_0$—that is, with $n = n_1 + n_2 + \cdots + n_k$ and $E_i = np_i$, where p_i is the probability of an observation being in the ith cell computed using the Poisson distribution with $\mu = \mu_0$. The p-value for the chi-square goodness-of-fit statistic is then obtained from Table 7 in the Appendix with $df = k - 1$, where k is the number of cells.

The second null hypothesis of interest to many researchers is less specific.

H_0: Data arise from a common Poisson model with μ unspecified.

H_a: Data do not arise from a Poisson model.

In this situation, it is necessary to first estimate μ using the data prior to computing an estimate of E_i. We then have $\hat{E}_i = n\hat{p}_i$, where $\hat{p}_i$s are obtained from a Poisson distribution with estimated parameter $\hat{\mu}$. The p-value for the chi-square goodness-of-fit statistic is then obtained from Table 7 in the Appendix with $df = k - 2$, where k is the number of cells. Note the difference in the degrees of freedom for the two measures of goodness-of-fit. For the null hypothesis with μ unspecified, it is necessary to reduce the degrees of freedom from $k - 1$ to $k - 2$ because we must first estimate the Poisson parameter μ prior to obtaining the cell probabilities.

For both types of hypotheses, we compute a p-value for the chi-square statistic and use this p-value to assess how well the model fits the data. Guidelines for assessing the quality of the fit are given here.

Guidelines for Assessing Quality of Model Fit

- p-value $\geq .25 \Rightarrow$ Excellent fit
- $.15 \leq p$-value $< .25 \Rightarrow$ Good fit
- $.05 \leq p$-value $< .15 \Rightarrow$ Moderately good fit
- $.01 \leq p$-value $< .05 \Rightarrow$ Poor fit
- p-value $< .01 \Rightarrow$ Unacceptable fit

The following example will illustrate the fit of a Poisson distribution to a data set.

EXAMPLE 10.13

Environmental engineers often utilize information contained in the number of different alga species and the number of cell clumps per species to measure the health of a lake. Those lakes exhibiting only a few species but many cell clumps are classified as oligotrophic. In one such investigation, a lake sample was analyzed under a microscope to determine the number of clumps of cells per microscope field. These data are summarized here for 150 fields examined under a microscope.

Here y_i denotes the number of cell clumps per field, and n_i denotes the number of fields with y_i cell clumps.

y_i	0	1	2	3	4	5	6	≥7
n_i	6	23	29	31	27	13	8	13

Use $\alpha = .05$ to test the null hypothesis that the sample data were drawn from a Poisson probability distribution.

Solution Before we can compute the value of χ^2, we must estimate the Poisson parameter μ and then compute the expected cell counts. The Poisson mean μ is estimated by using the sample mean $\bar{y}$. For these data,

$$\bar{y} = \frac{\sum_i n_i y_i}{n} = \frac{495}{150} = 3.3$$

Note that the sample mean was computed to be 3.3 by using all the sample data before the 13 largest values were collapsed into the final cell.

The Poisson probabilities for $y = 0, 1, \ldots, 7$ or more can be found in Table 14 in the Appendix with $\mu = 3.3$ or using the R function $dpois(x, 3.3)$, where $x = seq(0, 6, 1)$ and $P(y \geq 7) = 1 - ppois(6, 3.3)$. These probabilities are shown here.

y_i	0	1	2	3	4	5	6	≥ 7
$P(y_i)$ for $\mu = 3.3$	.0369	.1217	.2008	.2209	.1823	.1203	.0662	.0509

The expected cell count E_i can be computed for any cell using the formula $E_i = nP(y_i)$. Hence, for our data (with $n = 150$), the expected cell counts are as shown here.

y_i	0	1	2	3	4	5	6	≥7
E_i	5.54	18.26	30.12	33.14	27.35	18.05	9.93	7.63

Substituting these values into the test statistic, we have

$$\begin{aligned}\chi^2 &= \sum_i \left[\frac{(n_i - E_i)^2}{E_i}\right] \\ &= \frac{(6 - 5.54)^2}{5.54} + \frac{(23 - 18.26)^2}{18.26} + \cdots + \frac{(13 - 7.63)^2}{7.63} \\ &= 7.02 \text{ with df} = 8 - 2 = 6\end{aligned}$$

p-value $= Pr[\chi^2_6 > 7.02] = .319$ (using R function $1 - \textbf{\textit{pchisq}}(7.02, 6)$). Using Table 7 in the Appendix, we can conclude only that $.10 < p\text{-value} < .90$. Thus, using p-value $= .319$, we determine that the Poisson model provides an excellent fit to the data. ■

A word of caution is given here for situations in which we are considering this test procedure. As we mentioned previously, when using a chi-square statistic, we should have all expected cell counts fairly large. In particular, we want all $E_i > 1$ and not more than 20% less than 5. In Example 10.11, if values of $y \geq 7$ had been considered individually, the Es would not have satisfied the criteria for the use of χ^2. That is why we combined all values of $y \geq 7$ into one category.

The assumptions needed for running a chi-square goodness-of-fit test are those associated with a multinomial experiment, of which the key ones are independence of the trials and constant cell probabilities. Independence of the trials would be violated if, for example, several patients from the same family in Example 10.10 were included in the sample because hypertension has a strong

hereditary component. The assumption of constant cell probabilities would be violated if the study were conducted over a period of time during which the standards of medical practice shifted, allowing for other "standard" therapies.

The test statistic for the chi-square goodness-of-fit test is the sum of k terms, which is the reason the degrees of freedom depend on k, the number of categories, rather than on n, the total sample size. However, there are only $k - 1$ degrees of freedom, rather than k, because the sum of the $n_i - E_i$ terms must be equal to $n - n = 0$; $k - 1$ of the observed minus expected differences are free to vary, but the last one (kth) is determined by the condition that the sum of the $n_i - E_i$ equals zero.

This goodness-of-fit test has been used extensively over the years to test various scientific theories. Unlike previous statistical tests, however, the hypothesis of interest is the null hypothesis, not the research (or alternative) hypothesis. Unfortunately, the logic behind running a statistical test does not hold. In the standard situation in which the research (alternative) hypothesis is the one of interest to the scientist, we formulate a suitable null hypothesis and gather data to reject H_0 in favor of H_a. Thus, we "prove" H_a by contradicting H_0.

We cannot do the same with the chi-square goodness-of-fit test. If a scientist has a set theory and wants to show that sample data conform to or "fit" that theory, she wants to accept H_0. From our previous work, there is the potential for committing a Type II error in accepting H_0. Here, as with other tests, the calculation of β probabilities is difficult. In general, for a goodness-of-fit test, the potential for committing a Type II error is high if n is small or if k, the number of categories, is large. Even if the expected cell counts E_i conform to our recommendations, the probability of a Type II error could be large. Therefore, the results of a chi-square goodness-of-fit test should be viewed suspiciously. Don't automatically accept the null hypothesis as fact given that H_0 was not rejected.

10.5 Contingency Tables: Tests for Independence and Homogeneity

In Section 10.3, we showed a test for comparing two proportions. The data were simply counts of how many times we got a particular result in two samples. In this section, we extend that test. First, we present a single test statistic for testing whether several deviations of sample data from theoretical proportions could plausibly have occurred by chance.

When we first introduced probability ideas in Chapter 4, we started by using tables of frequencies (counts). At the time, we treated these counts as if they represented the whole population. In practice, we'll hardly ever know the complete population data; we'll usually have only a sample. When we have counts from a sample, they're usually arranged in **cross tabulations** or **contingency tables.** In this section, we'll describe one particular test that is often used for such tables, a chi-square test of independence.

cross tabulations
contingency tables

In Chapter 4, we introduced the idea of independence. In particular, we discussed the idea that **dependence** of variables means that one variable has some value for predicting the other. With sample data, there usually appears to be some degree of dependence. In this section, we develop a χ^2 test that assesses whether the perceived dependence in sample data may be a fluke—the result of random variability rather than real dependence.

dependence

First, the frequency data are to be arranged in a cross tabulation with r rows and c columns. The possible values of one variable determine the rows of the table,

and the possible values of the other determine the columns. We denote the population proportion (or probability) falling in row i, column j as π_{ij}. The total proportion for row i is $\pi_{i.}$ and the total proportion for column j is $\pi_{.j}$. If the row and column proportions (probabilities) are independent, then $\pi_{ij} = \pi_{i.}\pi_{.j}$.

For example, the Centers for Disease Control and Prevention wants to determine if the severity of a skin disease is related to the age of the patient. Suppose that a patient's skin disease is classified as moderate, mildly severe, or severe. The patients are divided into four age categories. Table 10.12 contains a set of proportions (π_{ij}) that exhibit independence between the severity of the disease and the age category in which the patient resides. That is, for each cell, $\pi_{ij} = \pi_{i.}\pi_{.j}$. For example, the proportion of patients who have a severe case of the disease and fall in age category I is $\pi_{31} = .02$. The proportion of all patients who have a severe case of the disease is $\pi_{3.} = .20$ and the proportion of all patients who fall in age category I is $\pi_{.1} = .10$. Independence holds for the (3,1) cell because $\pi_{31} = .02 = (.20)(.10) = \pi_{3.}\pi_{.1.}$ Similar calculations hold for the other eleven cells, and we can thus conclude that severity of the disease and age are independent.

TABLE 10.12
Distribution of skin disease over age categories

	Age Category				
Severity	**I**	**II**	**III**	**IV**	**All Ages**
Moderate	.05	.20	.15	.10	.50
Mildly severe	.03	.12	.09	.06	.30
Severe	.02	.08	.06	.04	.20
All severities	.10	.40	.30	.20	1.00

The null hypothesis for this χ^2 test is independence. The research hypothesis specifies only that there is some form of dependence—that is, that it is not true that $\pi_{ij} = \pi_{i.}\pi_{.j}$ in every cell of the table. The test statistic is once again the sum over all cells of

$$(\text{Observed value} - \text{expected value})^2/\text{expected value}$$

The computation of expected values E_{ij} under the null hypothesis is different for the independence test than for the goodness-of-fit test. The null hypothesis of independence does not specify numerical values for the row probabilities $\pi_{i.}$ and column probabilities $\pi_{.j}$, so these probabilities must be estimated by the row and column relative frequencies. If $n_{i.}$ is the actual frequency in row i, estimate $\pi_{i.}$ by $\hat{\pi}_{i.} = n_{i.}/n$; similarly, $\hat{\pi}_{.j} = n_{.j}/n$. Assuming the null hypothesis of independence is true, it follows that $\hat{\pi}_{ij} = \hat{\pi}_{i.}\hat{\pi}_{.j} = (n_{i.}/n)(n_{.j}/n)$.

DEFINITION 10.2

estimated expected value

Under the hypothesis of independence, the **estimated expected value** in row i, column j is

$$\hat{E}_{ij} = n\hat{\pi}_{ij} = n\frac{(n_{i.})}{n}\frac{(n_{.j})}{n} = \frac{(n_{i.})(n_{.j})}{n}$$

the row total multiplied by the column total divided by the grand total.

EXAMPLE 10.14

Suppose a random sample of 216 patients having the skin disease are classified into the four age categories, yielding the frequencies shown in Table 10.13.

TABLE 10.13 Results from random sample

	Age Category				
Severity	**I**	**II**	**III**	**IV**	**All Ages**
Moderate	15	32	18	5	70
Mildly severe	8	29	23	18	78
Severe	1	20	25	22	68
All severities	24	81	66	45	216

Calculate a table of $\hat{E}_{ij}$ values.

Solution For row 1, column 1, the estimated expected number of occurrences is

$$\hat{E}_{ij} = \frac{(\text{row 1 total})(\text{column 1 total})}{\text{grand total}} = \frac{(70)(24)}{216} = 7.78$$

Similar calculations for all cells yield the data shown in Table 10.14.

TABLE 10.14 Expected counts for example

	Age Category				
Severity	**I**	**II**	**III**	**IV**	**All Ages**
Moderate	7.78	26.25	21.39	14.58	70.00
Mildly severe	8.67	29.25	23.83	16.25	78.00
Severe	7.56	25.50	20.78	14.17	68.01
All severities	24.01	81.00	66.00	45.00	216.01

Note that the row and column totals in Table 10.13 equal (except for round-off error) the corresponding totals in Table 10.12. ■

χ^2 Test of Independence

H_0: The row and column variables are independent.
H_a: The row and column variables are dependent (associated).

T.S.: $\chi^2 = \sum_{i,j} (n_{ij} - \hat{E}_{ij})^2 / \hat{E}_{ij}$

R.R.: Reject H_0 if $\chi^2 > \chi^2_\alpha$, where χ^2_α cuts off area α in a χ^2 distribution with $(r-1)(c-1)$ df; r = number of rows, c = number of columns.

Check assumptions and draw conclusions.

The test statistic is referred to as the Pearson χ^2 statistic.

df for table

The **degrees of freedom** for the χ^2 test of independence relate to the number of cells in the two-way table that are free to vary while the marginal totals remain fixed. For example, in a 2×2 table (2 rows, 2 columns), only one cell entry is free

TABLE 10.14
(a) One df in a 2 × 2 table; (b) two df in a 2 × 3 table

	Category B		Total
Category A	*		16
			34
Total	21	29	50

(a)

	Category B			Total
Category A	*	*		51
				40
Total	28	41	22	91

(b)

to vary. Once that entry is fixed, we can determine the remaining cell entries by subtracting from the corresponding row or column total. In Table 10.14(a), we have indicated some (arbitrary) totals. The cell indicated by * could take any value (within the limits implied by the totals), but then all remaining cells would be determined by the totals. Similarly, with a 2 × 3 table (2 rows, 3 columns), two of the cell entries, as indicated by *, are free to vary. Once these entries are set, we determine the remaining cell entries by subtracting from the appropriate row or column total [see Table 10.14(b)]. In general, for a table with r rows and c columns, $(r-1)(c-1)$ of the cell entries are free to vary. This number represents the degrees of freedom for the χ^2 test of independence.

This chi-square test of independence is also based on an asymptotic approximation, which requires a reasonably large sample size. A conservative rule is that each $\hat{E}_{ij}$ must be at least 1 and no more than 20% of the $\hat{E}_{ij}$s can be less than 5 in order to obtain reasonably accurate p-values using the chi-square distribution. Standard practice when some of the $\hat{E}_{ij}$s are too small is to combine those rows (or columns) with small totals until the rule is satisfied. Care should be taken in deciding which rows (or columns) should be combined so that the new table is of an interpretable form. Alternatively, many software packages have an exact test that does not rely on the chi-square approximation.

EXAMPLE 10.15

Conduct a test to determine if the severity of the disease discussed in Example 10.14 is independent of the age of the patient. Use $\alpha = .05$, and obtain bounds on the p-value of the test statistic.

Solution The null and alternative hypotheses are

H_0: The severity of the disease is independent of the age of the patient.

H_a: The severity of the disease depends on the age of the patient.

The test statistic can be computed using the values of n_{ij} and $\hat{E}_{ij}$ from Example 10.12:

$$\text{T.S.:}\quad \chi^2 = \sum_{i,j}(n_{ij} - \hat{E}_{ij})^2/\hat{E}_{ij}$$
$$= (15 - 7.78)^2/7.78 + (32 - 26.25)^2/26.25$$
$$+ (18 - 21.39)^2/21.39 + \cdots + (22 - 14.17)^2/14.17$$
$$= 27.13$$

R. R.: For df $= (3-1)(4-1) = 6$ and $\alpha = .05$, the critical value from Table 7 in the Appendix is 12.59. Because $\chi^2 = 27.13$ exceeds 12.59, H_0 is rejected. The p-value $= Pr[\chi^2_6 > 27.13] = .00014$ using R. Based on the values in Table 7, we would conclude that p-value $< .001$.

Check the assumptions and draw conclusions: Since each of the estimated expected values $\hat{E}_{ij}$ exceeds 5, the chi-square approximation should be reasonably accurate. Thus, we can conclude that there is strong evidence in the data (p-value = .00014) that the severity of the disease is associated with the age of the patient. ■

likelihood ratio statistic

There is an alternative χ^2 statistic called the **likelihood ratio statistic** that is often shown in computer outputs. It is defined as

$$\text{likelihood ratio } \chi^2 = \sum_{ij} n_{ij} \ln(n_{ij}/(n_{i.}n_{.j}))$$

where $n_{i.}$ is the total frequency in row i, $n_{.j}$ is the total frequency in column j, and ln is the natural logarithm (base $e = 2.71828$). Its value should also be compared to the χ^2 distribution with the same $(r-1)(c-1)$ df. Although it isn't at all obvious, this form of the χ^2 independence test is approximately equal to the Pearson form. There is some reason to believe that the Pearson χ^2 yields a better approximation to table values, so we prefer to rely on it rather than on the likelihood ratio form.

The only function of a χ^2 test of independence is to determine whether apparent dependence in sample data may be a fluke, plausibly a result of random variation. Rejection of the null hypothesis indicates only that the apparent association is not reasonably attributable to chance. It does not indicate anything about the **strength** or type **of association.**

strength of association

The same χ^2 test statistic applies to a slightly different sampling procedure. An implicit assumption of our discussion surrounding the χ^2 test of independence is that the data result from a single random sample from the whole population. Often, separate random samples are taken from the subpopulations defined by the column (or row) variable. In the skin disease example (Example 10.12), the data might have resulted from separate samples (of respective sizes 24, 81, 66, and 45) from the four age categories rather than from a single random sample of 216 patients.

In general, suppose the column categories represent c distinct subpopulations. Random samples of size $n_1, n_2, \ldots, n_c$ are selected from these subpopulations. The observations from each subpopulation are then classified into the r values of a categorical variable represented by the r rows in the contingency table. The research hypothesis is that there is a difference in the distribution of subpopulation units into the r levels of the categorical variable. The null hypothesis is that the set of r proportions for each subpopulation $(\pi_{1j}, \pi_{2j}, \ldots, \pi_{rj})$ is the same for all $j = 1, 2, \ldots, c$ subpopulations. Thus, the null hypothesis is given by

$$H_0\colon (\pi_{11}, \pi_{21}, \ldots, \pi_{r1}) = (\pi_{12}, \pi_{22}, \ldots, \pi_{r2}) = \cdots = (\pi_{1c}, \pi_{2c}, \ldots, \pi_{rc})$$

test of homogeneity

The test is called a **test of homogeneity** of distributions. The mechanics of the test of homogeneity and the test of independence are identical. However, note that the sampling scheme and conclusions are different. With the test of independence, we randomly select n units from a single population and classify the units with respect to the values of two categorical variables. We then want to determine whether the two categorical variables are related to each other. In the test of homogeneity of proportions, we have c subpopulations from which we randomly select $n = n_1 + n_2 + \cdots + n_c$ units, which are classified according to the values of a single categorical variable. We want to determine whether the distribution of the subpopulation units to the values of the categorical variable is the same for all c subpopulations.

As we discussed in Section 10.4, the accuracy of the approximation of the sampling distribution of χ^2 by a chi-square distribution depends on both the sample size n and the number of cells k. Cochran (1954) indicates that the approximation should be adequate if no E_i is less than 1 and no more than 20% of the E_is are less than 5. Larntz (1978) and Koehler (1986) showed that χ^2 is valid with smaller sample sizes than is the likelihood ratio test statistic. Agresti (2002) compares the nominal and actual α-levels for both test statistics for testing independence, for various sample sizes. The χ^2 test statistic appears to be adequate when n/k exceeds 1. Again, we recommend applying Cochran's guidelines for determining whether the chi-square test statistic can be adequately approximated with a chi-square distribution. When some of the E_{ij}s are too small, there are several alternatives. Researchers combine levels of the categorical variables to increase the observed cell counts. However, combining categories should not be done unless there is a natural way to redefine the levels of the categorical variables that does not change the nature of the hypothesis to be tested. When it is not possible to obtain observed cell counts large enough to permit the chi-squared approximation, Agresti (2002) discusses *exact* methods to test the hypotheses. For example, the Fisher Exact test is used when both categorical variables have only two levels.

EXAMPLE 10.16

Random samples of 200 individuals from major oil-producing and natural gas–producing states, 200 from coal states, and 400 from other states participate in a poll of attitudes toward five possible energy policies. Each respondent indicates the most preferred alternative from among the following:

1. Primarily emphasize conservation
2. Primarily emphasize domestic oil and gas exploration
3. Primarily emphasize investment in solar-related energy
4. Primarily emphasize nuclear energy development and safety
5. Primarily reduce environmental restrictions and emphasize coal-burning activities

The results are as shown in Table 10.15.

TABLE 10.15
Results of survey

Policy Choice	Oil/Gas States	Coal States	Other States	Total
1	50	59	161	270
2	88	20	40	148
3	56	52	188	296
4	4	3	5	12
5	2	66	6	74
Total	200	200	400	800

Execustat output also carries out the calculations. The second entry in each cell is its percentage in the column.

```
                    Crosstabulation
                  OilGas     Coal    Other      Row
                                                Total
1                     50       59      161      270
                    25.0     29.5     40.3     33.75

2                     88       20       40      148
                    44.0     10.0     10.0     18.50

3                     56       52      188      296
                    28.0     26.0     47.0     37.00

4                      4        3        5       12
                     2.0      1.5      1.3      1.50

5                      2       66        6       74
                     1.0     33.0      1.5     9.25

Column               200      200      400      800
Total              25.00    25.00    50.00   100.00

    Summary Statistics for Crosstabulation

         Chi-square     D.F.    P Value

            289.22        8    0.0000

Warning: Some table cell counts < 5.
```

Conduct a χ^2 test of homogeneity of distributions for the three groups of states. Give the p-value for this test.

Solution A test that the corresponding population distributions are different makes use of the expected values found in Table 10.16.

TABLE 10.16
Expected counts for survey data

Policy Choice	Oil/Gas States	Coal States	Other States
1	67.5	67.5	135
2	37	37	74
3	74	74	148
4	3	3	6
5	18.5	18.5	37

We observe that the table of expected values has two E_{ij}s that are less than 5. However, our guideline for applying the chi-square approximation to the test statistic is met because only $2/15 = 13\%$ of the E_{ij}s are less than 5 and all the values are greater than 1. The test procedure is outlined here:

H_0: The column distributions are homogeneous.

H_a: The column distributions are not homogeneous.

T.S.: $\chi^2 = \sum (n_{ij} - \hat{E}_{ij})^2/\hat{E}_{ij}$

$= (50 - 67.5)^2/67.5 + (88 - 37)^2/37 + \cdots + (6 - 37)^2/37$

$= 289.22$

R.R.: Because the tabled value of χ^2 for df = 8 and $\alpha = .001$ is 26.12, p-value is $<.001$. Alternatively, use the R function p-value = 1 − ***pchisq***$(289.22, 8) \cong 0$ to many decimal places.

Check assumptions and draw conclusions: Even recognizing the limited accuracy of the χ^2 approximations, we can reject the hypothesis of homogeneity at some very small p-value. Percentage analysis, particularly of state type for a given policy choice, shows dramatic differences; for instance, 1% of those living in oil/gas states favor policy 5 compared to 33% of those in coal states who favor policy 5. ■

The χ^2 test described in this section has a limited but important purpose. This test assesses only whether the data indicate a statistically detectable (significant) relation among various categories. It does not measure how strong the apparent relation might be. A weak relation in a large data set may be detectable (significant); a strong relation in a small data set may be nonsignificant.

10.6 Measuring Strength of Relation

The χ^2 test we discussed in Section 10.5 has a built-in limitation. By design, the test answers only the question of whether there is a statistically detectable (significant) relation among the categories. It cannot answer the question of whether the relation is strong, interesting, or relevant. This is not a criticism of the test; no hypothesis test can answer these questions. In this section, we discuss methods for assessing the strength of relation shown in cross-tabulated data.

The simplest (and often the best) method for assessing the strength of a relation is simple percentage analysis. If there is no relation (that is, if complete independence holds), then percentages by row or by column show no relation. For example, suppose that a direct-mail company tests two different offers to see whether the response rates differ. Their results are shown in Table 10.17.

To check the relation, if any, we calculate percentages of response for each offer. We see that (40/200) =.20 (that is, 20%) respond to offer A and (80/400) = .20 respond to offer B. Because the percentages are exactly the same, there is no indication of relation. Alternatively, we note that one-third of the Yes respondents and one-third of the No respondents were given offer A. Because these fractions are exactly the same, there is no indication of a statistical relation.

Of course, it is rare to have data that show absolutely no relation in the sample. More commonly, the percentages by row or by column differ, suggesting some relation. For example, a firm planning to market a cleaning product commissions a market research study of the leading current product. The variables of interest are the frequency of use and the rating of the leading product. The data are shown in Table 10.18.

TABLE 10.17
Direct-mail responses

	Response		
Offer	**Yes**	**No**	**Total**
A	40	160	200
B	80	320	400
Total	120	480	600

TABLE 10.18
Responses from market survey

	Rating			
Use	**Fair**	**Good**	**Excellent**	**Total**
Rare	64	123	137	324
Occasional	131	256	129	516
Frequent	209	171	45	425
Total	404	550	311	1,265

To assess if there is a relationship between the level of use and the rating of the product by the consumer, we will first calculate the chi-square test of independence. We obtain $\chi^2 = 144.49$ with df $= (3-1)(3-1) = 4$. The p-value is computed as p-value $= Pr[\chi^2 > 144.49] < .001$, which would indicate strong evidence of a relationship between use and rating. The small p-value does *not* necessarily imply a strong relation; it could also be the result of a fairly weak relation but a very large sample size. We would next want to determine the type of relationship that may exist between use and rating. One natural analysis of these data takes the frequencies of use as given and looks at the ratings as functions of use. The analysis essentially looks at conditional probabilities of the rating factor given the level of the use factor. However, the analysis recognizes that the data are only a random sample, not the actual population values. For example, when the level of use is rare, the best estimate of the probability that the user will select a rating value of fair is determined using the formula

$$Pr\,[\text{Rating} = \text{Fair given User} = \text{Rare}] = \frac{64}{324} = .1975\ (19.75\%)$$

In a similar fashion, we compute

$$Pr\,[\text{Rating} = \text{Good given User} = \text{Rare}] = \frac{123}{324} = .3796$$

$$Pr\,[\text{Rating} = \text{Excellent given User} = \text{Rare}] = \frac{137}{324} = .4228$$

The corresponding proportions for occasional users are given by

$$Pr\,[\text{Rating} = \text{Fair given User} = \text{Occasional}] = \frac{131}{516} = .2539$$

$$Pr\,[\text{Rating} = \text{Good given User} = \text{Occasional}] = \frac{256}{516} = .4961$$

$$Pr\,[\text{Rating} = \text{Excellent given User} = \text{Occasional}] = \frac{129}{516} = .2500$$

For frequent users, the three proportions are

$$Pr\,[\text{Rating} = \text{Fair given User} = \text{Frequent}] = \frac{209}{425} = .4918$$

$$Pr\,[\text{Rating} = \text{Good given User} = \text{Frequent}] = \frac{171}{425} = .4024$$

$$Pr\,[\text{Rating} = \text{Excellent given User} = \text{Frequent}] = \frac{45}{425} = .1059$$

TABLE 10.19
Rating proportions from three types of users

	Rating		
Use	**Fair**	**Good**	**Excellent**
Rare	.1975	.3796	.4228
Occasional	.2539	.4961	.2500
Frequent	.4918	.4024	.1059

The proportions (or percentages, if one multiplies by 100) for the ratings are quite different for the three types of users, as can be seen in Table 10.19.

Thus, there appears to be a relation between the use variable and the ratings. The proportion of rare users giving the product an excellent rating is around 42%, whereas 25% of occasional users and only about 11% of frequent users give the product an excellent rating. Thus, as usage of the product increases the proportion of users giving an excellent rating decreases. The opposite is true for a rating of fair. The combination of a very small value for the *p*-value and a sizable difference in the conditional frequencies for the ratings depending on the level of usage provides substantial evidence that a relation between user and rating exists.

Percentage analyses play a fundamentally different role than does the χ^2 test. The point of a χ^2 test is to see how much evidence there is that there *is* a relation, whatever the size may be. The point of percentage analyses is to see *how strong* the relation appears to be, taking the data at face value. The two types of analyses are complementary.

Here are some final ideas about count data and relations:

1. A χ^2 goodness-of-fit test compares counts to theoretical probabilities that are specified outside the data. In contrast, a χ^2 independence test compares counts in one subset (one row, for example) to counts in other rows within the data. One way to decide which test is needed is to ask whether there is an externally stated set of theoretical probabilities. If so, the goodness-of-fit test is in order.
2. As is true of any significance test, the only purpose of a χ^2 test is to see whether differences in sample data might reasonably have arisen by chance alone. A test cannot tell you directly how large or important the difference is.
3. In particular, a statistically detectable (significant) χ^2 independence test does not necessarily mean a strong relation, nor does a nonsignificant goodness-of-fit test necessarily mean that the sample fractions are very close to the theoretical probabilities.
4. Looking thoughtfully at percentages is crucial in deciding whether the results show practical importance.

10.7 Odds and Odds Ratios

Another way to analyze count data on qualitative variables is to use the concept of odds. This approach is widely used in biomedical studies and could be useful in some market research contexts as well. The basic definition of odds is the ratio of the probability that an event happens to the probability that it does not happen.

DEFINITION 10.3

Odds of an event $A = \dfrac{P(A)}{1 - P(A)}$

If an event has probability 2/3 of happening, the odds are $\frac{2}{3}/\frac{1}{3} = 2$. Usually, this is reported as "the odds of the event happening are 2 to 1." Odds are used in horse racing and other betting establishments. The horse racing odds are given as the odds against the horse winning. Therefore, odds of 4 to 1 means that it is 4 times more likely the horse will lose (not win) than not. Based on the odds, a horse with 4 to 1 odds is a better "bet" than, say, a horse with 20 to 1 odds. What about a horse with 1 to 2 odds (or, equivalently, .5 to 1) against winning? This horse is highly favored because it is twice as likely (2 to 1) that the horse will win as not.

In working with odds, just make certain what the event of interest is. Also it is easy to convert the odds of an event back to the probability of the event. For event A,

$$P(A) = \frac{\text{odds of event } A}{1 + \text{odds of event } A}$$

Thus, if the odds of a horse (not winning) are stated as 9 to 1, then the probability of the horse not winning is

$$\text{Probability (not winning)} = \frac{9}{1 + 9} = .9$$

Similarly, the probability of winning is $1 - .9 = .1$.

Odds are a convenient way to see how the occurrence of a condition changes the probability of an event. Recall from Chapter 4 that the conditional probability of an event A given another event B is

$$P(A|B) = P(A \text{ and } B)/P(B)$$

The odds favoring an event A given another event B turn out after a little algebra to be

$$\frac{P(A|B)}{P(\text{not A}|B)} = \frac{P(A)}{P(\text{not } A)} \frac{P(B|A)}{P(B|\text{not } A)}$$

The initial odds are multiplied by the *likelihood ratio*, the ratio of the probability of the conditioning event given A to its probability given not A. If B is more likely to happen when A is true than when it is not, the occurrence of B makes the odds favoring A go up.

EXAMPLE 10.17

Consider both a population in which 1 of every 1,000 people carries the HIV virus and a test that yields positive results for 95% of those who carry the virus and (false) positive results for 2% of those who do not carry it. If a randomly chosen person obtains a positive test result, should the odds of that person carrying the HIV virus go up or go down? By how much?

Solution We certainly would think that a positive test result would increase the odds of carrying the virus. It would be a strange test indeed if a positive result

decreased the chance of having the disease! Take the event A to be "carries HIV" and the event B to be "positive test result."

Before the test is made, the odds of a randomly chosen person carrying HIV are

$$\frac{.001}{.999} \approx .001$$

The occurrence of a positive test result causes the odds to change to

$$\frac{P(\text{HIV}|\text{positive})}{P(\text{not HIV}|\text{positive})} = \frac{P(\text{HIV})}{P(\text{not HIV})} \frac{P(\text{positive}|\text{HIV})}{P(\text{positive}|\text{not HIV})} = \frac{.001}{.999} \frac{.95}{.02} = .0475$$

The odds of carrying HIV do go up given a positive test result—from about .001 (to 1) to about .0475 (to 1). ■

odds ratio

A closely related idea, widely used in biomedical studies, is the **odds ratio.** As the name indicates, it is the ratio of the odds of an event (for example, contracting a certain form of cancer) for one group (for example, men) to the odds of the same event for another group (for example, women). The odds ratio is usually defined using conditional probabilities but can be stated equally well in terms of joint probabilities.

DEFINITION 10.4

Odds Ratio of an Event for Two Groups

If A is any event with probabilities $P(A|\text{group 1})$ and $P(A|\text{group 2})$, the odds ratio (OR) is

$$OR = \frac{P(A|\text{group 1})/[1 - P(\text{A}|\text{group 1})]}{P(A|\text{group 2})/[1 - P(A|\text{group 2})]}$$

The odds ratio equals 1 if the event A is statistically independent of group.

We estimate the odds ratio in the following manner. Suppose we are investigating if there is a relation between the occurrence of a condition A and two groups. A random sample of n units is selected, and the number of units satisfying condition A are recorded for both groups, as displayed in Table 10.20.

The odds ratio compares the odds of the Yes proportion for group 1 to the odds of the Yes proportion for group 2. It is estimated from the observed data as

$$OR = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

TABLE 10.20
Data for computing an odds ratio

	Condition A			
	Yes	**No**	**Total**	**Proportion Yes**
Group 1	n_{11}	n_{12}	$n_{1.}$	$p_1 = n_{11}/n_{1.}$
Group 2	n_{21}	n_{22}	$n_{2.}$	$p_2 = n_{21}/n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n	

Inference about the odds ratio is usually done by way of the natural logarithm of the odds ratio. Recall that ln is the usual notation for the natural logarithm (base $e = 2.71828$) and that $\ln(1) = 0$. When the natural logarithm of the odds ratio is estimated from sampled data with a large value of n it has approximately a normal distribution with an expected value equal to the natural logarithm of the population odds ratio. Its standard error can be estimated by taking the square root of the sum of the reciprocals of the four counts in the above table.

Sampling Distribution of ln (OR)

For large sample sizes, the sampling distribution of the log odds ratio, ln(OR), is approximately normal with

$$\mu_{\ln(OR)} = \ln\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right)$$

where π_1 and π_2 are the population proportions for the two groups, and

$$\hat{\sigma}_{\ln(OR)} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

From the above results, we obtain an approximate $100(1-\alpha)$ confidence interval for the population log odds ratio, $\ln\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right)$:

$$(\ln(OR) - z_{\alpha/2}\hat{\sigma}_{\ln(OR)}, \ln(OR) + z_{\alpha/2}\hat{\sigma}_{\ln(OR)})$$

The above interval yields an approximate confidence interval for the population odds ratio by exponentiating the two endpoints of the interval. If this interval does *not* include an odds ratio 1.0, we conclude with $100(1-\alpha)$ confidence that there is substantial evidence that the event A is related to the groups.

EXAMPLE 10.18

A study was conducted to determine if the level of stress in a person's job affects his or her opinion about the company's proposed new health plan. A random sample of 3,000 employees yields the responses shown in Table 10.21.

TABLE 10.21
Relationship between job stress and health plan opinion

	Employee Response		
Job Stress	**Favorable**	**Unfavorable**	**Total**
Low	250	750	1,000
High	400	1,600	2,000
Total	650	2,350	3,000

Estimate the conditional probabilities of a favorable and an unfavorable response given the level of stress. Compute an estimate of the odds ratio of a favorable response for the two groups, and determine if type of response is related to level of stress.

Solution The estimated conditional probabilities are given in Table 10.22.

TABLE 10.22
Estimated conditional probabilities

	Employee Response		
Job Stress	**Favorable**	**Unfavorable**	**Total**
Low	.25	.75	1.0
High	.20	.80	1.0

The estimated odds ratio is $\frac{.25/.75}{.2/.8} = 1.333$. We could have computed the value of OR directly without having to first compute the conditional probabilities:

$$OR = \frac{(250)(1{,}600)}{(400)(750)} = 1.333$$

A value of 1.333 for the odds ratio indicates that the odds of a favorable response are 33.3% higher for employees in a low stress job than for employees with a high stress job. We will next compute a 95% confidence interval for the odds ratio and see if the confidence interval contains 1.0.

$$\ln(OR) = \ln(1.333) = 0.2874$$

and

$$\hat{\sigma}_{\ln(OR)} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = \sqrt{\frac{1}{250} + \frac{1}{750} + \frac{1}{400} + \frac{1}{1{,}600}} = \sqrt{.0084583}$$
$$= .0920$$

The 95% confidence interval for the odds ratio is obtained by first computing

$$(.2874 - (1.96)(0.0920), .2874 + (1.96)(0.0920)); \text{ that is, } (0.1071, 0.4677)$$

Exponentiating the endpoints then provides us with the confidence interval:

$$(e^{0.1071}, e^{0.4677}); \text{ that is, } (1.113, 1.5963)$$

Because the 95% confidence interval for the odds ratio does not include an odds ratio of 1.0, we may conclude that there is a statistically detectable relation between opinion and level of stress. ■

The odds ratio is a useful way to compare two population proportions π_1 and π_2 and may be more meaningful than their difference $(\pi_1 - \pi_1)$ when π_1 and π_2 are small. For example, suppose the rate of reinfarction for a sample of 5,000 coronary bypass patients treated with compound 1 is $\hat{\pi}_1 = .05$ and the corresponding rate for another sample of 5,000 coronary bypass patients treated with compound 2 is $\hat{\pi}_2 = .02$. Then their difference, $\hat{\pi}_1 - \hat{\pi}_2 = .03$, may be less important and less informative than the odds ratio. See Table 10.23.

TABLE 10.23
Reinfarction counts for bypass patients

	Reinfarction?		
	Yes	**No**	**Total**
Compound 1	250 (5%)	4,750	$n_1 = 5{,}000$
Compound 2	100 (2%)	4,900	$n_2 = 5{,}000$

The reinfarction odds for compounds 1 and 2 are as follows:

$$\text{Compound 1 odds} = \frac{250/5{,}000}{4{,}750/5{,}000} = \frac{250}{4{,}750} = .053$$

$$\text{Compound 2 odds} = \frac{100/5{,}000}{4{,}900/5{,}000} = \frac{100}{4{,}900} = .020$$

The corresponding odds ratio is .053/.020 = 2.65. Note that although the difference in reinfarction rates is only .033, having a reinfarction after treatment with compound 1 is 2.65 times as likely as having a reinfarction following treatment with compound 2.

10.8 Combining Sets of 2 × 2 Contingency Tables

In the previous section, we discussed the chi-square test of independence for examining the dependence of two variables based on data arranged in a contingency table. Suppose a pharmaceutical company is developing a drug product for the treatment of epilepsy. In each of several clinics, patients are assigned at random to either a placebo or the new drug and treated for a period of 2 months. At the end of the study, each patient is rated as either improved or not improved. If 100 patients (50 per treatment group) are to be enrolled in a particular clinic and we observe 40 and 15 patients improved in the new drug and placebo groups, respectively, the data could be displayed as shown in Table 10.24 and analyzed using the

TABLE 10.24
Number (%) of patients improved

	Improved	Not Improved	Total
New drug	40 (80%)	10	50
Placebo	15 (30%)	35	50

chi-square methods of the previous section. The null hypothesis of independence of the two classifications (treatment group and rating) could be restated in terms of the proportions, π_1 and π_2, of improved patients for the two populations. The new H_0 would be H_0: $\pi_1 - \pi_2 = 0$—namely, that there is no difference in the proportions of improved patients for the drug and placebo groups. Rejection of H_0 using the chi-square statistic from the test of independence test indicates that the population proportions are different for the two treatment groups.

This same scenario can be extended to more than one clinic, and we can extend our test procedure to deal with a set of q clinics ($q \geq 2$). For this situation, we would observe the sample percentages improved for the drug and placebo groups in each clinic; the data could be summarized using Table 10.25.

TABLE 10.25
Summary table for a set of 2 × 2 contingency tables

Clinic		Improved	Not Improved
1	Drug		
	Placebo		
2	Drug		
	Placebo		
	⋮		
q	Drug		
	Placebo		

TABLE 10.26
General notation for a set of 2 × 2 contingency tables

Table	Treatment	Response Category 1	Response Category 2	Total
1	1	n_{111}	n_{112}	$n_{11.}$
	2	n_{121}	n_{122}	$n_{12.}$
	Total	$n_{1.1}$	$n_{1.2}$	$n_{1..}$
2	1	n_{211}	n_{212}	$n_{21.}$
	2	n_{221}	n_{222}	$n_{22.}$
	Total	$n_{2.1}$	$n_{2.2}$	$n_{2..}$
⋮				
q	1	n_{q11}	n_{q12}	$n_{q1.}$
	2	n_{q21}	n_{q22}	$n_{q2.}$
	Total	$n_{q.1}$	$n_{q.2}$	$n_{q..}$

The test for comparing the drug and placebo proportions combines sample information across the separate contingency tables to answer the question of whether, on the average, the improvement rates are the same for the two treatment groups. Before we do this, however, we need some additional notation, shown in Table 10.26.

Cochran (1954) proposed a test statistic for the hypothesis of no difference (on the average) for the improvement rates for a set of q 2 × 2 contingency tables. This same problem was addressed by Mantel and Haenszel (1959) and also extended to cover a set of q 2 × c contingency tables. For 2 × 2 tables, the Cochran–Mantel–Haenszel (CMH) statistic for testing the equality of the improvement rates, on the average, can be written as

$$\chi^2_{\text{MH}} = \frac{\left\{\sum_h \left(n_{h11} - \frac{n_{h1.}n_{h.1}}{n_{h..}}\right)\right\}^2}{\sum_h \frac{n_{h1.}n_{h2.}n_{h.1}n_{h.2}}{n^2_{h..}(n_{h..} - 1)}}$$

which follows a chi-square distribution with df = 1. Let's see how this works for a set of sample data.

EXAMPLE 10.19

The pharmaceutical study discussed previously was extended to three clinics. In each clinic, as patients qualified for the study and gave their consent to participate, they were assigned to either the drug or the placebo group according to a predetermined random code. Each clinic was to treat 50 patients per group. The study results are summarized in Table 10.27. Use these data to test the null hypothesis of

TABLE 10.27
Study results

Clinic		Improved	Not Improved	Total
1	Drug	40 (80%)	10	50
	Placebo	15 (30%)	35	50
	Total	55	45	100
2	Drug	35 (70%)	15	50
	Placebo	20 (40%)	30	50
	Total	55	45	100
3	Drug	43 (86%)	7	50
	Placebo	31 (62%)	19	50
	Total	74	26	100
Total		184	116	300

no difference in the improvement rates, on the average. Use the CMH chi-square statistic, and give the p-value for the test.

Solution The necessary row and column totals in each clinic are given in Table 10.27. The numerator of the CMH statistic is

$$\left\{\sum_h \left(n_{h11} - \frac{n_{h1.}n_{h.1}}{n_{h..}}\right)\right\}^2 = \left\{\left(40 - \frac{50(55)}{100}\right) + \left(35 - \frac{50(55)}{100}\right) + \left(43 - \frac{50(74)}{100}\right)\right\}^2$$

$$= (12.5 + 7.5 + 6)^2 = 676$$

whereas the denominator is

$$\sum_h \frac{n_{h1.}n_{h2.}n_{h.1}n_{h.2}}{n_{h..}^2(n_{h..} - 1)} = \frac{50(50)(55)(45)}{(100)^2(99)} + \frac{50(50)(55)(45)}{(100)^2(99)} + \frac{50(50)(74)(26)}{(100)^2(99)}$$

$$= 6.25 + 6.25 + 4.8586 = 17.3586$$

Substituting, we obtain

$$\chi^2_{\text{MH}} = \frac{676}{17.3586} = 38.9432$$

For df = 1, this result is significant at the $p < .001$ level. As can be seen from the sample data, the drug-treated groups have consistently higher improvement rates than the placebo groups. ■

EXAMPLE 10.20

Sample data are not always as obvious and conclusive as those given in Example 10.19. Use the revised sample data shown in Table 10.28 to conduct a CMH test. Give the p-value for your test and interpret your findings.

TABLE 10.28
Revised study results

Clinic		Improved	Not Improved	Total
1	Drug	35 (70%)	15	50
	Placebo	26 (52%)	24	50
	Total	61	39	100
2	Drug	28 (56%)	22	50
	Placebo	29 (58%)	21	50
	Total	57	43	100
3	Drug	37 (74%)	13	50
	Placebo	24 (48%)	26	50
	Total	61	39	100

Solution Using the row and column totals of Table 10.28, the numerator and denominator of χ^2_{MH} can be shown to be 110.25 and 18.21, respectively. The CMH statistic is then

$$\chi^2_{MH} = 6.05$$

Based on df = 1, this test result has a significance defined by p-value = .0139. We conclude that although the drug product did not have a higher improvement rate in all three clinics, the data combined across clinics indicate that, on the average, there is significant evidence (p-value = .0139) that the drug improvement rate is higher than the placebo rate. ■

Mantel and Haenszel also extended this test procedure to cover the situation in which we want a combined test based on sample data displayed in a set of q $2 \times c$ contingency tables. Returning to our example, suppose rather than having two response categories (e.g., improved, not improved) we have c different categories (such as worse, same, slightly better, moderately better, completely well). For these situations, it is possible to score the categories of the scale and run a Mantel–Haenszel test based on the difference in mean scores for the two treatment groups. Because the formulas become more involved, available statistical software programs are used to make the calculations.

10.9 RESEARCH STUDY: Does Gender Bias Exist in the Selection of Students for Vocational Education?

In Section 10.1, we introduced some of the issues involved in gender bias.

Defining the Problem

The following questions would potentially be of interest to social scientists, civil rights advocates, and educators.

- Does gender play a role in the acceptance of a student into vocational education programs?
- What are some of the factors that may explain an association between gender and acceptance rate?

- How large a sample of students is needed to obtain substantial evidence of a bias or discrimination?

In this study, the researchers decided that they were initially interested in the overall acceptance and rejection rates for males and females in high school vocational education programs. To eliminate some of the potentially confounding factors, they decided to use only large public schools in northeastern states. In order to determine a sample size for the study, the researchers provided the following specifications: They wanted to be 95% confident that the estimated proportion of rejected applications was within .015 of the proportion of rejections in the population. Because school districts were reluctant to participate in the study, there was little insight with respect to what the population rejection rate would be. Thus, in calculating the sample size, a value of .50 (50%) was used. This yielded the following large-sample calculation:

$$n = \frac{(z^2_{.025})(.5)(1 - .5)}{(E)^2} = \frac{(1.96)(.5)(1 - .5)}{(.015)^2} = 4{,}268.4$$

It was decided to take a random sample of 5,000 students in order to obtain the desired degree of precision because a number of the students selected for the study might not have complete records.

Collecting the Data

A random sample of 1,000 applicants for vocational education was selected from each of five major northeastern school districts. Each of the 5,000 records provided the type of program that was applied for and whether the student was accepted or rejected for the program. The data were then summarized into tables and graphs.

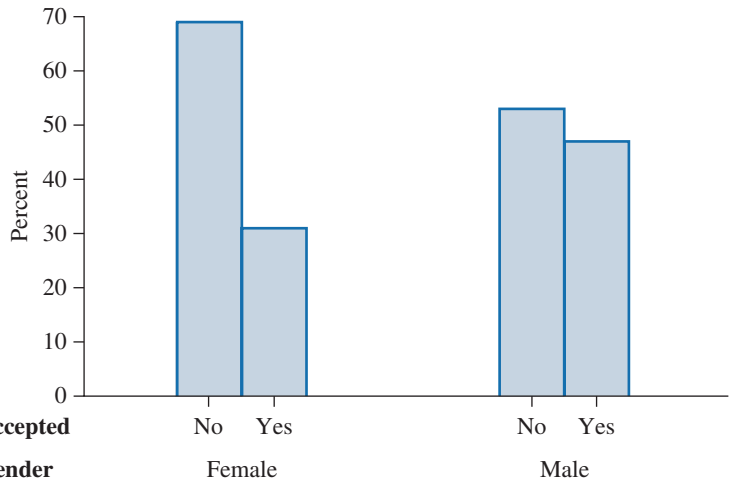
Summarizing the Data

Table 10.29 and Figure 10.1 summarize the data. A random sample of 5,000 high school students who have applied for vocational training is shown based on their gender and acceptance into the program. The cells contain the following information: count for each category, percentage of row, and percentage of column.

TABLE 10.29
Vocational training data

	Accepted in Program		
Gender	**No**	**Yes**	**All**
Female	963	433	1,396
	69.0%	31.0%	100.0%
	33.6%	20.3%	27.9%
Male	1,906	1,698	3,604
	52.9%	47.1%	100.0%
	66.4%	79.7%	72.1%
All	2,869	2,131	5,000
	57.4%	42.6%	100.0%

FIGURE 10.1
Acceptance percentages by gender



Analyzing the Data

From Figure 10.1, we can observe that female students have a much lower acceptance rate than do male students (31% versus 47.1%). To determine if this is a statistically significant difference, we test the following hypotheses:

H_0: Gender and acceptance are independent.

H_a: Gender and acceptance are associated.

Using a chi-square test of independence, we obtain $\chi^2 = 106.6$ with df = 1 and p-value $= Pr[\chi^2_1 > 106.6] < .0001$. Thus, there is strong evidence of an association between gender and acceptance into vocational education programs. To further explore this association, we note that the odds ratio of acceptance for males to acceptance for females is given by

$$OR = \frac{\text{male odds}}{\text{female odds}} = \frac{1{,}698/1{,}906}{433/963} = \frac{.8909}{.4496} = 1.98$$

with a 95% confidence interval of (1.67, 2.36). Thus, the odds of a male student being accepted into a vocational education program are nearly twice the odds of a female student. This is strong evidence of a bias in favor of male students.

The term bias is defined as an association between *the* acceptance or rejection decision and the gender of the applicant, which is very unlikely to have occurred just by chance. In order to validly use the odds ratio and chi-square tests of independence to support a conclusion of a bias, it is necessary for a couple of assumptions to hold. Bickel, Hammel, and O'Connell (1975) have a detailed discussion of these assumptions. Basically, assumption 1 is that male and female applicants for vocational education do not differ with respect to any attribute that is legitimately pertinent to their acceptance into a vocational education program. Assumption 2 is that the gender ratios of applicants to the various vocational education programs are not strongly associated with any other factors that are used in the acceptance decision methodology.

The researchers had decided to limit their study to only the four largest vocational education programs: plumbing, nursing, cosmetology, and welding. The aggregated data may be misleading due to the imbalance in the number of applicants by gender for the four programs. This could be a possible violation of assumption 2.

TABLE 10.30
Expanded vocational training data

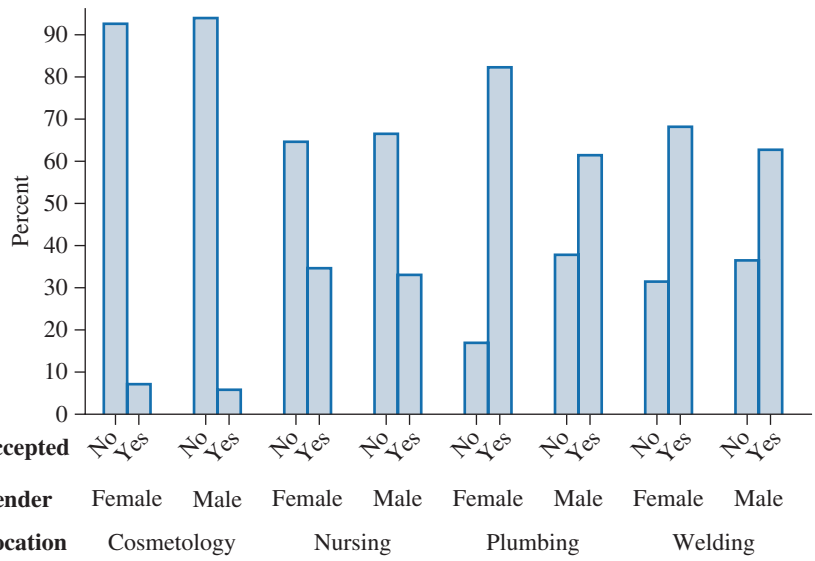
Vocation	Gender	Accepted	Frequency
Plumbing	Male	Yes	848
Welding	Male	Yes	585
Nursing	Male	Yes	229
Cosmetology	Male	Yes	36
Plumbing	Male	No	519
Welding	Male	No	343
Nursing	Male	No	462
Cosmetology	Male	No	582
Plumbing	Female	Yes	148
Welding	Female	Yes	28
Nursing	Female	Yes	217
Cosmetology	Female	Yes	40
Plumbing	Female	No	31
Welding	Female	No	13
Nursing	Female	No	404
Cosmetology	Female	No	515
All			5,000

That is, the gender ratios are associated with the type of vocational program. Table 10.30 and Figure 10.2 will examine the data separately for each of the four programs.

Figure 10.2 has consolidated the data across four major types of programs. Two of the programs are traditional male programs and two are traditional female programs. An analysis of the information about the types of programs the students applied for yields a more complete picture of the acceptance rates. The 5,000 applications are broken out by the type of vocational program applied for by the students.

Figure 10.2 displays the above data by plotting the percentages of acceptance and rejection within each level of gender and vocation. The pattern is much more

FIGURE 10.2
Acceptance rate by vocation and gender



complex than what was observed in Figure 10.1. In the aggregated data, females had a much lower acceptance rate than males (31.0% to 47.1%). However, when we examine the data by type of vocational program, we find that females have a higher percentage of acceptance than males in plumbing (82.7% versus 62.0%) and welding (68.3% versus 63.0%) with similar acceptance percentages in cosmetology (7.2% versus 5.8%) and nursing (34.9% versus 33.1%). These results appear to be impossible. Is this another case of deception through the manipulation of numbers by way of statistical methodology? There is no deception. This is an example of a lurking variable that confounds the association between gender and acceptance into the vocational education program. This type of data set has occurred often in the literature and is referred to as Simpson's Paradox.

The problem in the analysis of the aggregate data is that there is a violation of assumption 2. That is, the gender ratios are strongly associated with another factor that may be important in the study. In this study, the gender of the applicant is strongly associated with the type of vocational program. Table 10.31 displays the numbers and percentages of applicants by gender and type of program. The percentage of female applicants to the plumbing and welding programs is much lower than the corresponding percentages for males. A chi-square test of independence between the factors gender and type of program yields $\chi^2 = 940.3$ with df = 3 and p-value $< .0001$. Thus, there is strong evidence of an association between gender and type of vocational program. This association is the underlying factor that has distorted the results shown in the analysis of the aggregated data.

TABLE 10.31
Aggregated data for types of training

	Type of Program				
Gender	**Cosmetology**	**Nursing**	**Plumbing**	**Welding**	**All**
Female	555	621	179	41	1,396
	47.3%	47.3%	11.6%	4.2%	27.9%
Male	618	691	1,367	928	3,604
	52.7%	52.7%	88.4%	95.8%	72.1%
All	1,173	1,312	1,546	969	5,000

The data will now be analyzed separately for each of the four programs, and then an overall analysis using the Cochran–Mantel–Haenszel test statistic will be done. These results are summarized in Table 10.32.

Analyzing the Data Separately for Each Program

(a) Vocational Program—Cosmetology:

TABLE 10.32
Acceptance rates by gender and vocation program

	Accepted in Program		
Gender	**No**	**Yes**	**All**
Female	515	40	555
	92.8%	7.2%	100.0%
Male	582	36	618
	94.2%	5.8%	100.0%
All	1,097	76	1,173
	93.5%	6.5%	100.0%

a. $\chi^2 = .922$ with df = 1 and p-value = .337
b. $OR = .80$ with a 95% confidence interval of (.50, 1.27)

(b) Vocational Program—Nursing:

	Accepted in Program		
Gender	**No**	**Yes**	**All**
Female	404	217	621
	65.1%	34.9%	100.0%
Male	462	229	691
	66.9%	33.1%	100.0%
All	866	446	1,312
	66.0%	34.0%	100.0%

a. $\chi^2 = .474$ with df = 1 and p-value = .491
b. $OR = .92$ with a 95% confidence interval of (.73, 1.16)

(c) Vocational Program—Plumbing:

	Accepted in Program		
Gender	**No**	**Yes**	**All**
Female	31	148	179
	17.3%	82.7%	100.0%
Male	519	848	1,367
	38.0%	62.0%	100.0%
All	550	996	1,546
	35.6%	64.4%	100.0%

a. $\chi^2 = 29.44$ with df = 1 and p-value $< .0001$
b. $OR = .34$ with a 95% confidence interval of (.23, .51)

(d) Vocational Program—Welding:

	Accepted in Program		
Gender	**No**	**Yes**	**All**
Female	13	28	44
	31.7%	68.3%	100.0%
Male	343	585	928
	37.0%	63.0%	100.0%
All	356	613	969
	36.7%	63.3%	100.0%

a. $\chi^2 = .466$ with df = 1 and p-value = .495
b. $OR = .79$ with a 95% confidence interval of (.40, 1.55)

The Cochran–Mantel–Haenszel statistic with a continuity correction yields a value of 14.29 with a p-value = .00016. This would indicate that there is an association between gender and acceptance into a vocational education program. We can further analyze this association by examining each of the four programs individually. We observe that the confidence intervals for the odds ratios for three of the four programs contain 1.0. Only in the plumbing program does there appear to be a large difference in the acceptance rates for males and females. What can we conclude about a gender bias in the selection process for vocational education programs?

Communicating the Results

In the aggregate analysis, there was strong evidence that males had a much higher acceptance rate than females. When examining the four programs individually, the acceptance rate for females is higher than males in all four programs. This apparent contradiction occurs because there are large differences in the proportions of applicants by gender for the four programs. These differences would not have yielded such a large difference in the aggregate acceptance rate except for the fact that it was much more difficult for both genders to obtain acceptance in two of the programs (nursing and cosmetology). The overall acceptance rates were 34.0% for nursing and 6.5% for cosmetology, whereas the overall acceptance rates were 64.4% for plumbing and 63.3% for welding. This difference in acceptance rates is then magnified by the fact that the proportions of females who applied for admission were much lower than those of males in the programs having the higher acceptance rates. Thus, there appears to be a bias against female acceptance into vocational education programs when in fact females have higher acceptance rates in all four programs. When examining complex and socially difficult questions, it is very important that all factors of importance be included in the analysis in order to not reach an incorrect conclusion. Bickel, Hammel, and O'Connell (1975) provide much more in-depth analysis of this type of data.

10.10 Summary and Key Formulas

In this chapter, we dealt with categorical data. Categorical data on a single variable arise in a number of situations. We first examined estimation and test procedures for a population proportion (π) and for two population proportions ($\pi_1 - \pi_2$) based on independent samples. The extension of these procedures to comparing several population proportions (more than two) gave rise to the chi-square goodness-of-fit test.

Two-variable categorical data problems were discussed using the chi-square tests for independence and for homogeneity based on data displayed in an $r \times c$ contingency table. Fisher Exact test was introduced for analyzing 2×2 tables in which the expected counts are less than 5. The Cochran–Mantel–Haenszel test extends the chi-square test for independence to q sets of 2×2 tables.

Finally, we discussed odds and odds ratios, which are especially useful in biomedical trials involving binomial proportions.

Key Formulas

1. Confidence interval for π

$$\tilde{\pi} \pm z_{\alpha/2}\hat{\sigma}_{\tilde{\pi}}$$

where

$$\tilde{\pi} = \frac{\tilde{y}}{\tilde{n}}, \tilde{y} = y + .5z^2_{\alpha/2},$$

$$\tilde{n} = n + z^2_{\alpha/2}, \text{ and}$$

$$\hat{\sigma}_{\tilde{\pi}} = \sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{\tilde{n}}}$$

2. Sample size required for a $100(1-\alpha)\%$ confidence interval of the form $\hat{\pi} \pm E$

$$n = \frac{z^2_{\alpha/2}\pi(1-\pi)}{E^2}$$

(*Hint:* Use $\pi = .5$ if no estimate is available.)

3. Statistical test for π

$$\text{T.S.: } z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}}$$

where

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

4. Confidence interval for $\pi_1 - \pi_2$

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\,\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2}$$

where

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

5. Statistical test for $\pi_1 - \pi_2$

Case 1: Independent proportions $(\min(n_1\hat{\pi}_1, n_1(1-\hat{\pi}_1), n_2\hat{\pi}_2, n_2(1-\hat{\pi}_2)) \geq 5)$

T.S.: $z = \dfrac{\hat{\pi}_1 - \hat{\pi}_2}{\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2}}$

where

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

Case 2: Independent proportions $(\min(n_1\hat{\pi}_1, n_1(1-\hat{\pi}_1), n_2\hat{\pi}_2, n_2(1-\hat{\pi}_2)) < 5)$

T.S.: Fisher Exact test

Case 3: Correlated proportions $(n_{12} + n_{21} > 20)$

McNemar test

T.S.: $z = \dfrac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$

Case 4: Correlated proportions $(n_{12} + n_{21} \leq 20)$

T.S.: McNemar test—using binomial distribution

6. Multinomial distribution

$$P(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1!n_2!\ldots n_k!}\pi_1^{n}\pi_2^{n}\ldots\pi_k^{n}$$

7. Chi-square goodness-of-fit test

$$\text{T.S.: } \chi^2 = \sum_i \left[\frac{(n_i - E_i)^2}{E_i}\right]$$

where

$$E_i = n\pi_{i0}$$

8. Chi-square test of independence

$$\text{T.S.: } \chi^2 = \sum_{i,j} \left[\frac{(n_{ij} - E_{ij})^2}{E_{ij}}\right]$$

where

$$E_{ij} = \frac{(\text{row } i \text{ total})(\text{column } j \text{ total})}{n}$$

9. Odds of event $A = \dfrac{P(A)}{1-P(A)}$

$$\left(\text{in a binomial situation, odds of a success} = \frac{\pi}{(1-\pi)}\right)$$

10. Odds ratio for binomial situation, two groups

$$\frac{\text{Odds for group 1}}{\text{Odds for group 2}} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

11. Cochran–Mantel–Haenszel statistic

$$\chi^2_{\text{MH}} = \frac{\left\{\sum_h \left(n_{h11} - \dfrac{n_{h1.}\,n_{h.1}}{n_{h..}}\right)\right\}^2}{\sum_h \dfrac{n_{h1.}n_{h2.}\,n_{h.1}\,n_{h.2}}{n^2_{h..}(n_{h..}-1)}}$$

10.11 Exercises

10.2 Inferences About a Population Proportion π

Basic **10.1** For each of the following values for $\hat{\pi}$ and n, compute a 99% confidence interval for the population proportion π using both the standard large-sample procedure and the WAC adjusted procedure. Comment on whether the WAC adjustment was needed.

a. $n = 20, \hat{\pi} = .35$
b. $n = 35, \hat{\pi} = .80$
c. $n = 50, \hat{\pi} = .34$
d. $n = 100, \hat{\pi} = .12$

Basic **10.2** For each of the following values for $\hat{\pi}$ and n, compute a 95% confidence interval for the population proportion π using both the standard large-sample procedure and the WAC adjusted procedure. Comment on whether the WAC adjustment was needed.

a. $n = 20, \hat{\pi} = .35$
b. $n = 35, \hat{\pi} = .80$
c. $n = 50, \hat{\pi} = .34$
d. $n = 100, \hat{\pi} = .12$

Basic **10.3** For each of the following values for $\hat{\pi}$ and n, compute a 95% confidence interval for the population proportion π using both the standard large-sample procedure and the WAC adjusted procedure. Comment on whether the WAC adjustment was needed.

a. $n = 12, \hat{\pi} = .50$
b. $n = 25, \hat{\pi} = .20$
c. $n = 40, \hat{\pi} = .125$
d. $n = 100, \hat{\pi} = .05$

Basic **10.4** A random sample of 1,200 units is randomly selected from a population. If there are 732 successes in the 1,200 draws,

a. Construct a 95% confidence interval for π.
b. Construct a 99% confidence interval for π.
c. Explain the difference in the interpretation of the two confidence intervals.

Soc. **10.5** A public opinion polling agency plans to conduct a national survey to determine the proportion π of people who would be willing to pay a higher per kilowatt hour fee for their electricity provided the electricity was generated using ecologically friendly methods such as solar, wind, or nuclear. How many people must be included in the poll to estimate the population proportion within .04 of the population value using a 95% confidence interval. Consider two separate situations:

a. Suppose the polling agency has no previous information about the population proportion.
b. Suppose the polling agency is fairly certain that the population proportion is less than 30%.
c. Why are the sample sizes so different for the two situations?

Soc. **10.6** There has been a considerable underfunding of the national Highway Trust Fund over the last 20 years, which has resulted in a dramatic decline in the maintenance of the nation's roads and bridges. An organization representing highway construction companies is planning a nationwide survey to estimate the proportion π of people who would support an increase in the gasoline tax. How large a sample is needed to obtain an estimate of π to within .02 using a 99% confidence interval. Consider two separate situations:

a. Suppose the organization has no idea of the value of π.
b. Suppose the organization is fairly certain the value of π is greater than 75%.

Med. **10.7** The test for screening donated blood for the presence of the AIDS virus was developed in the 1980s. It is designed to detect antibodies, substances produced in the body of donors carrying the virus; however, it is not 100% accurate. The developer of the test claimed that the test would produce fewer than 5% false positives and fewer than 1% false negatives. In order to evaluate the accuracy

of the test, 1,000 persons known to have AIDS and 10,000 persons known to not have AIDS were given the test. The following results were tabulated:

	True State of Patient		
Test Result	**Has Aids**	**Does Not Have Aids**	**Total**
Positive test	993	591	1,584
Negative test	7	9,409	9,416
Total	1,000	10,000	11,000

a. Place a 99% confidence interval on the proportion of false positives produced by the test.

b. Is there substantial evidence ($\alpha = .01$) that the test produces less than 5% false positives.

Med. **10.8** Refer to Exercise 10.7.

a. Place a 99% confidence interval on the proportion of false negatives produced by the test.

b. Is there substantial evidence ($\alpha = .01$) that the test produces less than 2% false negatives.

c. Which of the two types of errors, false positives or false negatives, do you think is more crucial to public safety. Explain your reasoning.

Med. **10.9** Refer to Exercises 10.7 and 10.8. Although the accurate determination of the proportions of false positives and false negatives produced by an important medical test is important, the probabilities of the following two events are of greater interest. In the following two questions, you may assume that the point estimators of false positives and false negatives are the correct values of these probabilities. The prevalence of the AIDS virus in the population of people who donate blood is thought to be around 2%.

a. Suppose a person goes to a clinic and donates blood and the test of the AIDS virus results in a positive test result. What is the probability that the person donating blood actually is carrying the AIDS virus?

b. Suppose a person goes to a clinic and donates blood and the test of the AIDS virus results in a negative test result. What is the probability that the person donating blood does not have the AIDS virus?

Med. **10.10** In a study of self-medication practices, a random sample of 1,230 adults completed a survey. The survey reported that 441 of the persons had a cough or cold during the past month and 260 of these individuals said they had treated the cough or cold with an over-the-counter (OTC) remedy. The data are summarized next.

Respondents reporting cough or cold	441
Respondents using an OTC remedy	260
Respondents using specific class of OTC remedy:	
Pain relievers	110
Cold capsules	57
Cough remedies	44
Allergy remedies	9
Liquid cold remedies	35
Nasal sprays	4
Cough drops	13
Sore-throat lozenges	9
Room vaporizers	4
Chest rubs	9

a. Provide a graphical display of the above data using percentages. Do your percentages add to 100%? Why or why not?
b. Based on the above data, for what classes of OTC remedies could you validly obtain a 95% confidence interval for the corresponding population proportion π?

Edu. **10.11** An administrator at a university with an average enrollment of 55,000 students wants to estimate the proportion of students who would support an increase in the student activity fee. This increase would be used to fund a $450 million renovation of the campus football stadium. How many students would need to be selected if the administrator wants to be 99% confident that the sample estimator is within .05 of the proportion for the whole campus.

Bio. **10.12** An entomologist is studying a new tick species that may be the carrier of the pathogen associated with lyme disease. She designs a study to estimate prevalence of the pathogen in the tick. She examines 100 ticks randomly selected in the study region during a period of the year when ticks have been known to be infected with the pathogen in other regions of the country. The examination of the 100 ticks finds none of the ticks are infected with the pathogen.

a. Provide the entomologist with an estimate of the proportion of ticks of this species that are carrying the pathogen.
b. Construct a 95% confidence interval for the proportion of ticks of this species that are carrying the pathogen.
c. The prevalence of the lyme-associated pathogen in the black-legged tick is 2%. Is the prevalence in the new species of tick less than the prevalence in the black-legged tick? Use $\alpha = .01$.

Bus. **10.13** The sales manager for a very exclusive brand of automobile declares that, after his staff had completed a new training program, less than 10% of their clients were dissatisfied with the service obtained at the dealership. The owner of the dealership hires a marketing firm to evaluate this claim. In a random sample of 40 customers, only 5 of the 40 customers were dissatisfied with the dealerships service.

a. Is there significant evidence that the sales manager's claim is supported by the data? Use α.05.
b. Place a 95% confidence interval on the proportion of customers who are dissatisfied with the service in their encounters with the staff at the dealership.

Med. **10.14** Chronic pain is often defined as pain that occurs constantly and flares up frequently, is not caused by cancer, and is experienced at least once a month for a 1-year period of time. Many articles have been written about the relationship between chronic pain and the age of the patient. In a survey conducted on behalf of ***the American Chronic Pain Association in 2004***, a random cross section of 800 adults who suffer from chronic pain found that 424 of the 800 participants in the survey were above the age of 50.

a. Would it be appropriate to use a normal approximation in conducting a statistical test of the research hypothesis that over half of persons suffering from chronic pain are over 50 years of age?
b. Using the data in the survey, is there substantial evidence ($\alpha = .05$) that more than half of persons suffering from chronic pain are over 50 years of age?
c. Place a 95% confidence interval on the proportion of persons suffering from chronic pain that are over 50 years of age.

Pol. Sci. **10.15** National public opinion polls are often based on as few as 1,500 persons in a random sampling of public sentiment on issues of public interest. These surveys are often done in person because the response rate for a mailed survey is very low and telephone interviews tend to reach a larger proportion of older persons than would be represented in the public as a whole. Suppose a random sample of 1,500 registered voters was surveyed about energy issues.

a. If 230 of the 1,500 responded that they would favor drilling for oil in national parks, estimate the proportion π of registered voters who would favor drilling for oil in national parks. Use a 95% confidence interval.
b. How many persons must the survey include to have 95% confidence that the sample proportion is within .01 of π?
c. A congressman has claimed that over half of all registered voters would support drilling in national parks. Use the survey data to evaluate the congressman's claim. Use $\alpha = .05$.

10.3 Inferences About the Difference Between Two Population Proportions, $\pi_1 - \pi_2$

Basic **10.16** A random sample of 250 observations is taken from population A, which has 30% of its population living in poverty: $\pi_A = .3$. A second random sample of size 350 is taken independently taken from population B, which has 15% of its population living in poverty: $\pi_B = .15$.

- **a.** What are the mean and standard deviation of the difference in the sample proportions, $\hat{\pi}_A - \hat{\pi}_B$?
- **b.** Describe the shape of the sampling distribution of the difference in the sample proportions, $\hat{\pi}_A - \hat{\pi}_B$?
- **c.** Is it appropriate to use the normal approximation to the sampling distribution of the difference in the sample proportions, $\hat{\pi}_A - \hat{\pi}_B$?

Basic **10.17** Refer to Exercise 10.16. Assuming that equal sample sizes will be taken from the two populations, how large a sample should be taken from each of the populations to obtain a 99% confidence interval for $\pi_A - \pi_B$ with a width of at most .02? (*Hint:* Use $\hat{\pi}_A = .3$ and $\hat{\pi}_B = .15$ from Exercise 10.16.

Bus. **10.18** A large retail lawn care dealer currently provides a 2-year warranty on all lawn mowers sold at its stores. A new employee suggested that the dealer could save money by just not offering the warranty. To evaluate this suggestion, the dealer randomly decides whether or not to offer the warranty to the next 50 customers who enter the store and express an interest in purchasing a lawnmower. Out of the 25 customers offered the warranty, 10 purchased a mower as compared to 4 of 25 not offered the warranty.

- **a.** Place a 95% confidence interval on $\pi_1 - \pi_2$, the difference in the proportions of customers purchasing lawnmowers with and without the warranty.
- **b.** Test the research hypothesis that offering the warranty will increase the proportion of customers who will purchase a mower. Use $\alpha = .05$.
- **c.** Are the conditions for using a large-sample test to answer the question in part (b) satisfied? If not, apply an exact procedure.
- **d.** Based on your results from parts (a) and (b), should the dealer offer the warranty?

Bus. **10.19** An advertising agency is considering two advertisements for a major client. One of the advertisements is in black and white, and the other is in color. A market research firm randomly selects 50 male and 50 female customers of the client to evaluate the two advertisements. The firm finds that 39 of the 50 males prefer the color advertisement, whereas 46 of the 50 females preferred the color advertisement.

- **a.** Place a 95% confidence on the difference in the proportions of males and females that prefer the color advertisement.
- **b.** Does the confidence interval indicate that there is a significant difference in the proportions? Use $\alpha = .05$.
- **c.** Are the conditions for using a large-sample test to answer the question in part (b) satisfied? If not, apply an exact procedure.
- **d.** Based on your results from parts (a) and (b), should the advertisement firm use different advertisements for male and female customers?

Med. **10.20** Biofeedback is a treatment technique in which people are trained to improve their health by using signals from their own bodies. Specialists in many different fields use biofeedback to help their patients cope with pain. A study was conducted to compare a biofeedback treatment for chronic pain with an NSAID medical treatment. A group of 2,000 newly diagnosed chronic pain patients were randomly assigned to receive to one of the two treatments. After 6 weeks of treatments, the pain levels of the patients were assessed with the following results:

	Significant Reduction in Pain		
Treatment	**Yes**	**No**	**Total**
Biofeedback	560	440	1,000
NSAID	680	320	1,000
Total	1,240	760	2,000

a. For both treatments, place 95% confidence intervals on the proportions of patients who experienced a significant reduction in pain.
b. Is there significant evidence ($\alpha = .05$) of a difference in the two treatments relative to the proportions of patients who experienced a significant reduction in pain?
c. Place a 95% confidence interval on the difference in the two proportions.

Ag. **10.21** Sludge is a dried product remaining from processed sewage and is often used as a fertilizer on agriculture crops. If the sludge contains a high concentration of certain heavy metals, such as nickel, the nickel may reach such a concentration in the crops that it becomes a danger to the consumer of the crops. A new method of processing sewage has been developed, and an experiment is conducted to evaluate its effectiveness in removing heavy metals. Sewage of a known concentration of nickel is treated using both the new and the old methods. One hundred tomato plants were randomly assigned to pots containing sewage sludge processed by one of the two methods. The tomatoes harvested from the plants were evaluated to determine if the nickel was at a toxic level. The results are as follows:

	Level of Nickel		
Treatment	**Toxic**	**Nontoxic**	**Total**
New	5	45	50
Old	9	41	50
Total	14	86	100

a. For both treatments, place 95% confidence intervals on the proportions of plants that would have a toxic level of nickel.
b. Is there significant evidence ($\alpha = .05$) that the new treatment would produce a lower proportion of plants having a toxic level of nickel compared to the old treatment?
c. Use the Fisher Exact test to test the research hypothesis that the new treatment would produce a lower proportion of plants having a toxic level of nickel compared to the old treatment. Compare your conclusions with the conclusions reached in part (b).
d. Place a 95% confidence interval on the difference in the two proportions.

Pol. Sci. **10.22** A political scientist is studying the impact of a political debate between candidates for governor in a small western state. The scientist wants to evaluate the proportion of registered votes who switch their preference after viewing the debate. The following table contain the data from 75 registered voters.

Preference	**Preference After Debate**	
Before Debate	**Candidate A**	**Candidate B**
Candidate A	28	13
Candidate B	6	28

a. Test whether there was a shift away from candidate A after the debate. Use $\alpha = .05$. Carefully state your conclusion.
b. Construct a 95% confidence interval for the change after the debate in the proportion of registered voters preferring candidate A.

Engin. **10.23** An article by ***Chen and Chen (1995)*** compared the quality of two speech recognition systems. A benchmark of 2000 words was submitted to both the generalized minimal distortion segmentation (GMDS) system and the continuous density hidden Markov model (CDHMM). The following table contains the results of the performance of the two systems, the number of words correctly and incorrectly identified.

	CDHMM	
GMDS	**Correct**	**Incorrect**
Correct	1,921	58
Incorrect	16	5

a. Are the proportions of correct identifications by the two systems correlated or independent? Justify your answer.
b. Test whether the proportions of correct identifications differ for the two systems. Use $\alpha = .05$. Carefully state your conclusion in terms of the population parameters.
c. Construct a 95% confidence interval for the difference in the two systems proportions of correct identifications.

Soc. **10.24** The question of whether the sexual orientation of the mother has an impact on the sexual identity of her children was addressed in an article by ***Golombok and Tasker (1996)***. Twenty-five children of lesbian mothers and a control group of 21 children of heterosexual single mothers were interviewed in their early twenties concerning their sexual orientation. The results of the interviews are given in the following table. Data were unavailable for 1 male child; thus, orientation is reported on only 20 children of heterosexual single mothers.

	Child's Orientation	
Mother's Orientation	**Nonheterosexual**	**Heterosexual**
Lesbian	2	23
Heterosexual	0	20

a. What is the populations of interest in this study?
b. Is the proportion of young adults identifying themselves as nonheterosexual higher for lesbian mothers than for heterosexual mothers? Use $\alpha = .05$.
c. Construct a 95% confidence interval on the difference in the two proportions of young adults who identify themselves as being heterosexual in their sexual orientation.

10.4 Inferences About Several Proportions: Chi-Square Goodness-of-Fit Test

Basic **10.25** List the characteristics of a multinomial experiment.

Basic **10.26** How does a binomial experiment relate to a multinomial experiment?

Basic **10.27** Under what conditions is it appropriate to use the chi-square goodness-of-fit test for the proportions in a multinomial experiment? What qualification(s) might one have to make if the sample data do not yield a rejection of the null hypothesis?

Basic **10.28** What restrictions are placed on the sample size n in order to appropriately apply the chi-square goodness-of-fit test?

Bus. **10.29** The quality control department of a motorcycle company classifies new motorcycles according to the number of defective components per motorcycle at an initial inspection. An improvement to the production process has been implemented, and, hopefully, there will be a change from the historical defective distribution: $\pi_1 = .80$, $\pi_2 = .10$, $\pi_3 = .05$, $\pi_4 = .03$, and $\pi_5 = .02$. A random sample of 300 motorcycles produced under the new system is classified as follows:

Number of Defectives	Number of Motorcycles, n_i
0	238
1	32
2	12
3	13
4 or more	5
Total	300

At the $\alpha = .05$ level, does there appear to be a change in the historical proportions of defectives?

Bus. **10.30** Refer to Exercise 10.29.

a. Place 95% confidence intervals on the proportions of the production falling into the five classifications.
b. Do the confidence intervals support the conclusion reached in Exercise 10.29?
c. Why may the conclusion reached using the confidence intervals differ from the conclusion reached in Exercise 10.29?

Soc. **10.31** The data in the following table from the book ***A Handbook of Small Data Sets (Hand et al., 1993, p. 36)*** document the starting positions of the winning horses in 144 races. The starting position listed as 1 is the position of the horse in the starting gate closest to the inside rail of the track, and position 8 is farthest from the rail. Racing officials contend that starting position has no effect on the chance of winning the race.

Starting position	1	2	3	4	5	6	7	8
Number of winners	29	19	18	25	17	10	15	11

a. What is the population of interest?
b. Do the data support the racing officials' contention?

Soc. **10.32** The article ***"Positive Aspects of Caregiving" (Research on Aging (2004) 26:429–453)*** describes a study that assessed how caregiving to Alzheimer's patients impacted the caregivers. Most people would generally think that family members who provide daily care to parents and spouses with Alzheimer's disease would tend to be negatively impacted by their role as caregiver. The study asked 1,229 caregivers to respond to the following statement: "Caregiving enabled me to develop a more positive attitude toward life." The following responses were reported:

	Response					
	Disagree a Lot	**Disagree a Little**	**No Opinion**	**Agree a Little**	**Agree a Lot**	**Total**
Number	166	116	171	234	542	1,229
% of total	13.5	9.4	13.9	19.2	44.1	100

a. Provide a graphical display of the data that illustrates potential differences in the percentages in the five cells.
b. Is there significant evidence that the proportions are not equally dispersed over the five possible responses? Use $\alpha = .05$.
c. Based on the graph in part (a) and your conclusions from part (b), does providing care to Alzheimer's patients have generally a positive or negative impact on caregivers?

Soc. **10.33** Organizations interested in making sure that accused persons have a trial of their peers often compare the distribution of jurors by age, education, and other socioeconomic variables. One such study in a large southern county provided the following information on the ages of 1,000 jurors and the age distribution countywide.

	Age				
	21–40	**41–50**	**51–60**	**Over 60**	**Total**
Number of jurors	399	231	158	212	1,000
Age % countywide	42.1	22.9	15.7	19.3	100

a. Display the above data using appropriate graphs.
b. Is this significant evidence of a difference between the age distribution of jurors and the countywide age distribution?
c. Does there appear to be an age bias in the selection of jurors?

Soc. **10.34** Refer to Exercise 10.33. The following information displays the education distribution of 1,000 jurors and the education distribution countywide.

	Education Level				
	Elementary	**Secondary**	**College Credits**	**College Degree**	**Total**
Number of jurors	278	523	98	101	1,000
Education % countywide	39.2	40.5	9.1	11.2	100

a. Display the above data using appropriate graphs.
b. Is this significant evidence of a difference between the education distribution of jurors and the countywide education distribution?
c. Does there appear to be bias in the selection of jurors with respect to the education level of jurors?

Bus. **10.35** A researcher obtained a sample of 125 security analysts and asked each analyst to select four stocks on the New York Stock Exchange that were expected to outperform the Standard and Poor's Index over a 3-month period. One theory suggests that the securities analysts would be expected to do no better than chance. Hence, the number of correct guesses from the four selected stocks for any analyst would have a binomial distribution with $n = 4$ and $\pi = .5$ yield probabilities, as shown here:

	Number Outperforming				
	0	**1**	**2**	**3**	**4**
Multinomial probabilities (π_i)	.0625	.25	.375	.25	.0625

The number of analysts' selections that outperformed the Standard and Poor's Index are given here:

	Number Outperforming					
	0	**1**	**2**	**3**	**4**	**Total**
Frequency	3	23	51	39	9	125

Do the data support the contention that the analysts' performance is different from just randomly selecting four stocks?

Hist. **10.36** A study examining bomb hits in South London during World War II is documented in the following table from the book ***A Handbook of Small Data Sets (Hand et al., 1993, p. 232).*** The bomb hits were recorded in the 576 grids in a map of a region in South London. The study contended that certain areas were less likely to be hit with a bomb because of certain geographical features. If the bomb hits were purely random, a Poisson model would produce the number of hits per grid.

Number of bomb hits	0	1	2	3	4	5	6	7	Total
Number of grids	229	211	93	35	7	0	0	1	576

a. Does the distribution of bomb hits appear to be random across this region of South London?
b. State the null and research hypotheses for this study in terms of π_i, the probability of i bomb hits in a grid.

Engin. **10.37** Nylon bars were tested for brittleness. Each of 280 bars was molded under similar conditions and was tested by placing a fixed stress at specified locations on the bar. Assuming that each bar has uniform composition, the number of breaks on a given bar should be Poisson distributed with an unknown a rate of breaks, λ, appearing per square inch of bar. The following table summarizes the number of breaks found on the 280 bars:

Breaks/bar	0	1	2	3	4	5	Total
Frequency	121	110	38	7	3	1	280

a. Use a goodness-of-fit test to assess whether the data appear to be from a Poisson model.

b. What is the population of interest in this study?

Bio. **10.38** A genetics experiment on the characteristics of tomato plants provided the following data on the number of offspring expressing four phenotypes.

Phenotype	Tall, cut-leaf	Dwarf, cut-leaf	Tall, potato-leaf	Dwarf, potato-leaf	Total
Frequency	926	293	288	104	1,611

a. The researcher wants to determine if there is substantial evidence that the tomato plants deviate from the current theory that the four phenotypes will appear in the proportion 9:3:3:1. Use $\alpha = .05$.

b. What is the population of interest in this study?

Bio. **10.39** Entomologists study the distribution of insects across agricultural fields. A study of fire ant hills across pasture lands is conducted by dividing pastures into 50-meter squares and counting the number of fire ant hills in each square. The null hypothesis of a Poisson distribution for the counts is equivalent to a random distribution of the fire ant hills over the pasture. Rejection of the hypothesis of randomness may occur due to one of two possible alternatives. The distribution of fire ant hills may be uniform—that is, the same number of hills per 50-meter square—or the distribution of fire ants may be clustered across the pasture. A random distribution would have the variance in counts equal to the mean count, $\sigma^2 = \mu$. If the distribution is more uniform than random, then the distribution is said to be underdispersed, $\sigma^2 < \mu$. If the distribution is more clustered than random, then the distribution is said to be overdispersed, $\sigma^2 > \mu$. The number of fire ant hills was recorded on one hundred 50-meter squares. In the data set, y_i is the number of fire ant hills per square, and n_i denotes the number of 50-meter squares with y_i ant hills.

y_i	0	1	2	3	4	5	6	7	8	9	12	15	20
n_i	2	6	8	10	12	15	13	12	10	6	3	2	1

a. Estimate the mean and variance of the number of fire ant hills per 50-meter square; that is, compute $\bar{y}$ and s^2 using the formulas from Chapter 3.

b. Do the fire ant hills appear to be randomly distributed across the pastures? Use a chi-square test of the adequacy of the Poisson distribution to fit the data using $\alpha = .05$.

c. If you reject the Poisson distribution as a model for the distribution of fire ant hills, does it appear that fire ant hills are more clustered or uniformly distributed across the pastures?

10.5 Contingency Tables: Tests for Independence and Homogeneity

H.R. **10.40** The recruitment director for a large engineering firm categorizes universities based on their rankings by *U.S. News* as most desirable, desirable, adequate, or undesirable for purposes of hiring the engineering graduates from the universities. The director reviews the performance records of 156 engineers employed by the firm for 1–2 years. The following table cross-classifies the annual performance ratings of the engineers with the universities from which they earned their BS degrees.

	Performance Rating of Employee		
University Type	**Outstanding**	**Average**	**Poor**
Most desirable	21	20	4
Desirable	4	26	36
Adequate	13	7	2
Undesirable	10	7	6

a. What is the population of interest represented by the data in the above table?
b. Can the director conclude that there is a relationship between the university type and the performance rating of the employee?
c. Are the conditions for applying your test in part (b) satisfied?

H.R. **10.41** Refer to Exercise 10.40. Suppose the recruitment director is preparing a presentation for upper management to recommend new hiring practices of the firm.

a. Provide a graph of the data in Exercise 10.40.
b. Comment on the results from Exercise 10.40 in terms of do whether hiring practices over the past 1–2 years appear to be successful. If not, suggest some changes.

Gov. **10.42** The fire department in a large city is examining its promotion policy to assess if there is the potential for an age discrimination lawsuit. A random sample of 248 promotion decisions over the past 5 years yields the following information.

	Age at Promotion Decision			
Promotion Decision	**Under 30**	**30–39**	**40–49**	**50 or Older**
Promoted	9	29	34	12
Not promoted	41	39	46	38

a. Provide a graph of the promotion information.
b. Is the promotion decision for the fireman related to the age of the fireman? Use $\alpha = .05$.
c. What is the population to which your conclusion in part (b) is applicable?
d. What are some other variables, besides age, that needed to be addressed in an age discrimination analysis?

Gov. **10.43** Refer to Exercise 10.42. Suppose that the initial analysis of the age discrimination had included only two levels of age, as contained in the following table.

	Age at Promotion Decision	
Promotion Decision	**39 or younger**	**40 or Older**
Promoted	38	46
Not promoted	80	84

a. Is the age of the fireman related to whether or not the fireman is promoted? Use $\alpha = .05$.
b. Is your conclusion concerning age discrimination different from your conclusion using the data in Exercise 10.42?

Ag. **10.44** Integrated Pest Management (IPM) adopters apply significantly less insecticides and fungicides than nonadopters among grape producers. The paper ***"Environmental and Economic Consequences of Technology Adoption: IPM in Viticulture" [Agricultural Economics (2008) 18:145–155]*** contained the following adoption rates for the six states that account for most of the U.S. production. A survey of 712 grape-producing growers asked whether or not the growers were using an IPM program on the farms.

	State						
	Cal.	**Mich.**	**New York**	**Oregon**	**Penn.**	**Wash.**	**Total**
IPM adopted	39	55	19	22	24	30	189
IPM not adopted	92	69	114	88	83	77	523
Total	131	124	133	110	107	107	712

a. Provide a graphical display of the data.
b. Is there significant evidence that the proportions of grape farmers who have adopted IPM are different across the six states?

Ag. **10.45** Refer to Exercise 10.44. Suppose that the grape farmers in the states of California, Michigan, and Washington were provided with information about the effectiveness of IPM by the county agents, whereas the farmers in the remaining states were not.

a. Is there significant evidence that providing information about IPM is associated with a higher adoption rate?

b. Discuss why or why not your conclusion in part (b) provides justification for expanding the program for county agents to discuss IPM with grape farmers to other states.

Soc. **10.46** Social scientists have produced convincing evidence that parental divorce is negatively associated with the educational success of their children. The paper ***"Maternal Cohabitation and Educational Success" [Sociology of Education (2005) 78:144–164]*** describes a study that addresses the impact of cohabiting mothers on the success of their children in graduating from high school. The following table displays the educational outcome by type of family for 1,168 children.

	Type of Family					
	Two-Parent	Single-Parent		Stepparent		
High Schl. Grad.		Always	Divorce	No Cohab.	With Cohab.	Total
Yes	407	61	231	124	193	1,016
No	45	16	29	11	51	152
Total	452	77	260	135	244	1,168

a. Display the above data in a graph to demonstrate any differences in the proportions of high school graduates across family types.

b. Is there significant evidence that the proportions of students who graduate from high school are different across the various family types?

Soc. **10.47** Refer to Exercise 10.46. For those students living within a stepparent family, does cohabitation appear to affect high school graduation rates?

10.6 Measuring Strength of Relation

H.R. **10.48** Refer to Exercise 10.40. Provide a description of the relationship between the four types of universities and the performance ratings of the newly hired engineers.

Gov. **10.49** Refer to Exercise 10.42. Describe the relationship between the ages of the firemen at promotion and the promotion decisions.

10.50 Refer to Exercise 10.44. Describe the type of relationship that exists between the various states and the proportions of farms at which an IPM program was adopted.

10.51 Refer to Exercise 10.46. Describe the type of relationship that exists between the family types and the proportions of students who graduated from high school.

10.7 Odds and Odds Ratios

Med. **10.52** A food-frequency questionnaire is used to measure dietary intake. The respondent specifies the number of servings of various food items he or she consumed over the previous week. The dietary cholesterol is then quantified for each respondent. The researchers were interested in assessing if there was an association between dietary cholesterol intake and high blood pressure. In a large sample of individuals who had completed the questionnaire, 250 persons with a high dietary cholesterol intake (greater than 300 mg/day) were selected, and 250 persons with a low dietary cholesterol intake (less than 300 mg/day) were selected. The 500 selected participants had their medical histories taken and were classified as having normal or high blood pressure. The data are given here.

	Blood Pressure		
Dietary Cholesterol	**High**	**Low**	**Total**
High	159	91	250
Low	78	172	250
Total	237	263	500

a. Compute the difference in the estimated risks of having high blood pressure $(\hat{\pi}_1 - \hat{\pi}_2)$ for the two groups (low versus high dietary cholesterol intake).
b. Compute the estimated relative risks of having high blood pressure $\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right)$ for the two groups (low versus high dietary cholesterol intake).
c. Compute the estimated odds ratio of having high blood pressure for the two groups (low versus high dietary cholesterol intake).
d. Based on your results from parts (a)–(c), how do the two groups compare?

H.R. **10.53** Refer to Exercise 10.52.
a. Compare the low and high dietary cholesterol intake groups relative to their risks of having high blood pressure. Use $\alpha = .05$.
b. Place a 95% confidence interval on the odds ratio of having high blood pressure for low cholesterol intake to having high blood pressure for high cholesterol intake. Based on the confidence interval, what can you conclude about the odds of having high blood pressure for the two groups?
c. Are your conclusions from parts (a) and (b) consistent?

Safety **10.54** The article ***"Who Wants Airbags" [Chance (2005 18:3–16]*** discusses whether air bags should be mandatory equipment in all new automobiles. From National Highway Traffic Safety Administration (NHTSA) data, the authors obtained the following information about fatalities and the usage of air bags and seat belts. All passenger cars sold in the United States starting in 1998 are required to have air bags. The NHTSA estimates that air bags had saved 10,000 lives as of January 2004. The authors examined accidents in which there was a harmful event (personal or property) and from which at least one vehicle was towed. After some screening of the data, they obtained the following results. (The authors detail in their article the types of screening of the data that was done.)

	Air Bag Installed		
	Yes	**No**	**Total**
Killed	19,276	27,924	47,200
Survived	5,723,539	4,826,982	10,550,521
Total	5,742,815	4,854,906	10,597,721

a. Calculate the odds of being killed in a harmful-event car accident for vehicles with and without air bags. Interpret the two odds.
b. Calculate the odds ratio of being killed in a harmful-event car accident with and without air bags. What does this ratio tell you about the importance of having air bags in a vehicle?
c. Is there significant evidence of a difference between vehicles with and without air bags relative to the proportions of persons killed in harmful-event vehicle accidents? Use $\alpha = .05$.
d. Place a 95% confidence interval on the odds ratio. Interpret this interval.

10.55 Refer to Exercise 10.54. The authors also collected information about accidents concerning seat belt usage. The article compared fatality rates for occupants using seat belts properly with those for occupants not using seat belts. The data are given here.

	Seat Belt Usage		
	Seat Belt	**No Seat Belt**	**Total**
Killed	16,001	31,199	47,200
Survived	7,758,634	2,791,887	10,550,521
Total	7,774,635	2,823,086	10,597,721

a. Calculate the odds of being killed in a harmful-event car accident for vehicle occupants who were using seat belts and those who were not using seat belts. Interpret the two odds.
b. Calculate the odds ratio of being killed in a harmful-event car accident with and without seat belts being used properly. What does this ratio tell you about the importance of using seat belts?
c. Is there significant evidence of a difference between vehicles with and without proper seat belt usage relative to the proportions of persons killed in a harmful-event vehicle accident? Use $\alpha = .05$.
d. Place a 95% confidence interval on the odds ratio. Interpret this interval.

10.56 Refer to Exercises 10.54 and 10.55. Which of the two safety devices appears to be more effective in preventing a death during an accident? Justify your answer using the information from the previous two exercises.

10.57 Refer to Exercises 10.54 and 10.55. To obtain a more accurate picture of the impact of air bags on preventing deaths, it is necessary to account for the effect of occupants using both seat belts and air bags. If the occupants of the vehicles in which air bags are installed are more likely to be also wearing seat belts, then it is possible that some of the apparent effectiveness of the air bags is in fact due to the increased usage of seat belts. Thus, one more 2×2 table is necessary: the table displaying a comparison of proper seat belt usage for occupants with air bags available and for occupants without air bags available. Those data are given here.

	Seat Belt Usage		
Air Bags	**Seat Belt**	**No Seat Belt**	**Total**
Yes	4,871,940	870,875	5,742,815
No	2,902,694	1,952,211	4,854,905
Total	7,774,634	2,823,086	10,597,720

a. Is there significant evidence of an association between air bag installation and the proper usage of seat belts? Use $\alpha = .05$
b. Provide justification for your results in part (a).

10.58 With reference to the information provided in Exercises 10.54, 10.55, and 10.57, there was one more question of interest to the researchers. If people in cars with air bags are more likely to be wearing seat belts, then how much of the improvement in fatality rates with air bags is really due to seat belt usage? The harmful-event fatalities were then classified according to both availability of air bags and seat belt usage. The data are given here.

	Seat Belt Usage		
Air Bags	**Seat Belt**	**No Seat Belt**	**Total**
Yes	8,626	10,650	19,276
No	7,374	20,550	27,924
Total	16,000	31,200	47,200

a. Use the information in the previous table and the data from Exercise 10.57 to compute the fatality rates for the four air bag and seat belt combinations.
b. Describe the confounding effect of seat belt usage on the effect of air bags on reducing fatalities.

Supplementary Exercises

Engin. **10.59** The police department supplies its officers with a flashlight that contains four batteries. The company that manufacturers the flashlights is required to verify the reliability of the batteries that are included in the flashlight when it is shipped to the police department. The quality control department states that at most 15% of its batteries are defective. The four batteries were inspected in a random sample of 300 flashlights, and the numbers of defective batteries are listed in the following table.

Number of defective batteries	0	1	2	3	4	Total
Frequency	100	126	60	13	1	300

a. Estimate π, the probability that a battery is defective. (*Hint:* What is the total number of batteries in the 300 flashlights?)
b. Place a 95% confidence interval on π.
c. Is there strong evidence to refute the claim that at most 15% of its batteries are defective?

Engin. **10.60** Refer to Exercise 10.59.
a. Does a binomial model with $n = 4$ and $\pi = .15$, where π is the probability that an individual battery is defective, appear to fit the above data? (*Hint:* Let $D =$ number of defective batteries in a flashlight:
$\pi_k = P(D = k) = \frac{4}{k!(4-k)!}(.15)^k(.85)^{4-k} =$ ***dbinom***(k, 4, .15))
b. Let $\hat{\pi}$ be the estimate of π from Exercise 10.59. Does a binomial model with $n = 4$ and $\hat{\pi}$ appear to fit the above data?

10.61 Another study from the book ***A Handbook of Small Data Sets (Hand et al., 1993)*** describes the family structure in the Hutterite Brethren, a religious group that is essentially a closed population with nearly all marriages involving members of the group. The researchers were interested in studying the offsprings of such families. The following data list the distribution of sons in families with seven children.

	Number of Sons							
	0	1	2	3	4	5	6	7
Frequency	0	6	14	25	21	22	9	1

a. Test the hypothesis that the number of sons in a family of seven children follows a binomial distribution with $\pi = .5$. Use $\alpha = .05$.
b. Suppose that π is unspecified. Evaluate the general fit of a binomial distribution. Using the p-value from your test statistic, comment on the adequacy of using a binomial model for this situation.
c. Compare your results from parts (a) and (b).

10.62 An entomologist was interested in determining if Colorado potato beetles were randomly distributed over a potato field or if they tended to appear in clusters. The field was gridded into evenly spaced squares, and counts of the beetle were conducted. The following data give the number of squares in which various numbers of beetles were observed. If the appearance of the potato beetle is random, a Poisson model should provide a good fit to the data.

	Number of Beetles						
	0	1	2	3	4	5 or more	Total
Number of squares	678	227	56	28	8	14	1,011

a. The average number of beetles per square is 0.5. Does the Poisson distribution provide a good fit to the data?
b. Based on your results in part (a), do Colorado potato beetles appear randomly across the field?

Ag. **10.63** A retail computer dealer is trying to decide between two methods for servicing customers' equipment. The first method emphasizes preventive maintenance; the second emphasizes quick response to problems. The dealer serves samples of customers by one of the two methods in a random fashion. After 6 months, the dealer finds that 171 of 200 customers serviced by the first method are very satisfied with the service as compared to 155 of 200 customers served by the second method.

a. Test the research hypothesis that the population proportions of very satisfied customers are different for the two methods. Use $\alpha = .05$. Carefully state your conclusion.
b. Compute a confidence interval for the difference in the proportions. Does the confidence interval provide the same conclusion about the difference in proportions as your test in part (a)? Justify your answer.

Engin. **10.64** To evaluate the difference in the reliabilities of cooling motors for PCs from two suppliers, an accelerated life test is performed on 50 motors randomly selected from the warehouses of the two suppliers. Supplier A's motors are considerably more expensive in comparison to the motors of supplier B. Of the motors from supplier A, 37 were still running at the end of the test period, whereas only 27 of the 50 motors from supplier B were still running at the end of the test period.

a. Is there significant evidence that supplier A's motors are more reliable than supplier B's motors? Use $\alpha = .05$.
b. Use the Fisher Exact test to test the research hypothesis that supplier A's motors are more reliable than supplier B's motors. Compare your conclusion with the conclusion reached in part (a).
c. Calculate 95% confidence intervals for the proportions of motors that passed the test for each supplier and for the difference in the two proportions. Interpret the results carefully in terms of the reliability of the two suppliers' motors.

Bio. **10.65** A research entomologist is interested in evaluating a new chemical formulation for possible use as a pesticide for controlling fire ants. She decides to compare its performance relative to the most widely used pesticide on the market, AntKiller. Each of the pesticides is applied to 100 containers of fire ants. The new pesticide successfully killed all the fire ants within 2 hours of application in 65 of the 100 containers. Of the 100 containers treated with AntKiller, only 59 had all fire ants killed.

a. Is there significant evidence that the proportion of containers successfully treated by the new formulation is greater than the proportion successfully treated by AntKiller? Use $\alpha = .05$.
b. Use the Fisher Exact test to test the research hypothesis that the proportion of containers successfully treated by the new formulation is greater than the proportion successfully treated by AntKiller? Use $\alpha = .05$. Compare your conclusion to the conclusion reached in part (a).
c. Place a 95% confidence interval on the difference in the two proportions.
d. Based on the results in parts (a)–(c), can the entomologist claim that she has shown that the new formulation is more effective than AntKiller?

Ag. **10.66** A new treatment is developed for controlling aphid infestation in sorghum. In order to assess the effectiveness of the treatment, 100 sorghum plants were treated, and 100 sorghum plants were left untreated. The 200 plants were then exposed to aphids, and 1 week later the plants were classified into three categories of infestation, as given in the following table.

	Level of Infestation		
Treatment	**Leaf Infestation**	**Stem Infestation**	**No Infestation**
Treated	11	29	60
Control	37	39	24

a. Provide a graph to compare the difference in infestations between the treatment and control.
b. Do the levels of infestation differ between the treatment and control plants? Use $\alpha = .05$.
c. What are the populations to which your conclusion in part (b) is applicable?
d. Place 95% confidence intervals on the probabilities of infestation for both the treatment and the control plants.

Bus. **10.67** Three different television commercials are advertising an established product. The commercials are shown separately to theater panels of consumers; each consumer views only one of the possible commercials and then states an opinion of the product. Opinions range from 1 (very favorable) to 5 (very unfavorable). The data are as follows.

	Opinion					
Commercial	**1**	**2**	**3**	**4**	**5**	**Total**
A	32	87	91	46	44	300
B	53	141	76	20	10	300
C	41	93	67	36	63	300
Total	126	321	234	102	117	900

a. Calculate expected frequencies under the null hypothesis of independence.
b. How many degrees of freedom are available for testing this hypothesis?
c. Is there evidence that the opinion distributions are different for the various commercials? Use $\alpha = .01$.

Bus. **10.68** Refer to Exercise 10.67 Provide a description of the relationship between the three types of commercials and the opinions of panels of consumers.

Ag. **10.69** A study was conducted to compare two anesthetic drugs for use in minor surgery using 45 men who were similar in age and physical condition. The two drugs were applied on the right and left ankles of each patient, and after a fixed period of time, the doctor recorded whether or not the ankle remained anesthetized. Data from the 45 patients are recorded below:

	Drug 2 Response	
Drug 1 Response	**Remains Anesthetized**	**Not Anesthetized**
Remains Anesthetized	12	10
Not Anesthetized	9	14

a. Is there significant evidence of a difference in the effectiveness of the two drugs. Use $\alpha = .05$.
b. Place a 95% confidence on the difference in the effectiveness of the two drugs.

Med. **10.70** A study reported in ***Meehan et al. (2013)*** was conducted to determine whether athletes had sustained previous undiagnosed concussions. A total of 731 patients met the inclusion criteria and were enrolled during the study period. Of these, 227 patients (31%) were unable to answer the questions that were used to determine if they had a previously undiagnosed concussion. An additional 18 were removed for incomplete or inaccurate data. Thus, 486 patients were included in the final analysis with a mean age of 15.5 years (with a standard deviation of 3.5 years). Most participants (63%) were male. The athletes playing a given sport at the time of the current injury were then classified according to the sport they participated in and whether or not they had a previously undiagnosed concussion.

	Sport of Current Injury				
Previous Unreported Concussion	**Football**	**Ice Hockey**	**Soccer**	**Basketball**	**Lacrosse**
Yes	32	25	20	13	7
No	66	58	49	31	24

a. Is the proportion of athletes having a previously unreported concussion related to the sport in which they participated?
b. What are the populations to which your conclusion in part (a) is applicable?

Med. **10.71** Refer to Exercise 10.70. Provide a description of the relationship between the sport the athlete was participating in and whether the athlete had sustained a previously unreported concussion.

Bus. **10.72** A legal software firm has created a more comprehensive but also more complex version of the software used by counties to manage their court systems. The company selected a few current customers to beta test the new software. Each of the persons who used the old software was evaluated using a survey and then assigned a rating to reflect his or her level of sophistication in terms of using software. The ratings are basic user, moderately complex user, and highly complex user. After using the new software for a few weeks, the individual users then responded with a level of preference of the new software compared to the current version of the software. The levels were strong preference for current version, moderate preference for current version, no preference, moderate preference for new version, and strong preference for new version. The data for the 190 current users of the software are given in the following table.

Sophistication of User	Preference of User				
	Strong Curr.	Mod. Curr.	No Prefer.	Mod. New	Strong New
Basic	32	28	17	12	4
Moderate complex	10	16	20	10	8
Highly complex	2	4	5	8	14

a. Is there evidence of a significant relationship between sophistication of the user and level of preference? Use $\alpha = .05$.
b. Describe the relationship between sophistication of the user and level of preference (if any).

Bus. **10.73** A large used-book store randomly selected 1,000 of its customers and asked them to complete a survey about their satisfaction with the the merchandise in the store. The following data are from the 224 customers who returned the survey. Although the survey requested a considerable amount of information, the store was most interested in the frequency of purchases (number of books purchased in past 3 months) and the customer's rating of the adequacy of book selection in the store. The data from the 224 surveys are given in the following table.

Frequency of Purchases	Adequacy of Selection			
	Poor	Average	Good	Excellent
1	3	4	37	44
2	2	6	30	28
3	3	8	16	19
4 or more	2	12	5	5

a. Is there evidence of a significant relationship between frequency of purchases and adequacy of selection? Use $\alpha = .05$.
b. Describe the relationship between frequency of purchases and adequacy of selection (if any).

Bus. **10.74** Refer to Exercise 10.73.
a. Were the conditions necessary to perform the test in Exercise 10.73 satisfied?
b. If the conditions were not satisfied, perform an alternative analysis by combining some of the categories.
c. Compare the conclusion obtained from the test in Exercise 10.73 to the conclusion from part (b).
d. What are some of the issues with using your conclusions when only 21.5% of the customers responded to the survey?

Bus. **10.75** Refer to Exercise 10.73. For each of the four levels of frequency of purchase, compute the proportion of customers in each of the four adequacy of selection categories. Describe the trends in the proportions across the adequacy of selection categories. What differences do you see in the four trends?

Bus. **10.76** A major bank surveyed a random sample of 398 employees to determine whether they preferred having an HMO or a traditional fee-for-service medical benefit plan. The survey categorized the employees by age and medical plan preference, with the outcomes given below.

No Dependents Covered

Age of Employee	Medical Plan Preference				
	Strong HMO	Modest HMO	Neutral	Modest For Fee	Strong For Fee
20–29	13	17	8	2	1
30–39	6	11	3	2	3
40–49	5	2	3	1	1
50–59	4	5	1	0	2
60 or older	5	3	2	3	2

1 or More Dependents Covered

Age of Employee	Medical Plan Preference				
	Strong HMO	Modest HMO	Neutral	Modest For Fee	Strong For Fee
20–29	3	0	3	7	3
30–39	5	9	10	22	21
40–49	13	6	21	24	25
50–59	1	7	11	9	13
60 or older	4	1	52	8	15

Combine the two groups of employees and answer the following questions.

- **a.** Is there evidence of a significant relationship between the age of the employee and medical plan preference? Use $\alpha = .05$.
- **b.** Describe the relationship between the age of the employee and medical plan preference (if any).

Bus. **10.77** Refer to Exercise 10.76. There may be an indirect relation between employee age and medical plan preference: Age might be related to whether an employee has dependents covered, and whether dependents are covered might be related to medical plan preference.

- **a.** Is there evidence of a significant relationship between the age of the employee and whether the employee has dependents covered by a plan? Use $\alpha = .05$.
- **b.** Is there evidence of a significant relationship between strength of preference for a medical plan and whether the employee has dependents covered by a plan? Use $\alpha = .05$.
- **c.** Finally, separately for the two groups of employees, test if there is evidence of a significant relationship between the age of the employee and preference for a medical plan? Use $\alpha = .05$.
- **d.** Based on the analyses in parts (a)–(c), what are your conclusions about the relationships among the age of the employee, medical plan preference, and whether dependents are covered by a plan?

Bio. **10.78** A carcinogenicity study was conducted to examine the tumor potential of a drug product scheduled for initial testing in humans. A total of 300 rats (150 males and 150 females) was studied for a 6-month period. At the beginning of the study, 100 rats (50 males, 50 females) were randomly assigned to the control group, 100 (50 males, 50 females) to the low-dose group, and the remaining 100 (50 males, 50 females) to the high-dose group. On each day of the 6-month

period, the rats in the control group received an injection of an inert solution, whereas those in the drug groups received an injection of the solution plus drug. The sample data are shown in the accompanying table.

	Number of Tumors	
Rat Group	**One or More**	**None**
Control	10	90
Low dose	14	86
High dose	19	81

a. Conduct a test of whether there is a significant difference in the proportions of rats having one or more tumors for the three treatment groups with $\alpha = .05$.
b. Does there appear to be a drug-related problem regarding tumors for this drug product? That is, as the dose is increased, does there appear to be an increase in the proportion of rats with tumors?

Bus. **10.79** Refer to Exercise 10.78.
a. Compare the odds of a tumor appearing for each of the three rat groups.
b. Place a 95% confidence interval on the odds ratio of a tumor appearing for the control group to appearing for the low-dose group.
c. Place a 95% confidence interval on the odds ratio of a tumor appearing for the control group to appearing for the high-dose group.
d. Place a 95% confidence interval on the odds ratio of a tumor appearing for the low-dose group to appearing for the high-dose group.
e. What are your conclusions about the impact of the drug product on tumor appearance?

Soc. **10.80** A sociological study was conducted to determine whether there is a relationship between the length of time blue-collar workers remain in their first job and the amount of their education. From union membership records, a random sample of persons was classified. The data are shown here.

	Years of Education			
Years on First Job	**0–4**	**5–9**	**10–12**	**13 or more**
0–2	5	21	30	33
3–5	15	35	40	30
6–8	22	16	15	30
9 or more	28	10	8	10

a. Test the research hypothesis that the variable "years on first job" is related to the variable "years of education."
b. Give the level of significance for the test.
c. Draw your conclusions using $\alpha = .05$.

Psy. **10.81** Two researchers at Johns Hopkins University studied the use of drug products in the elderly. Patients in a recent study were asked the extent to which physicians counseled them with regard to their drug therapies. The researchers found the following:

- 25.4% of the patients said their physicians did not explain what the drug was supposed to do.
- 91.6% indicated they were not told how the drug might "bother" them.
- 47.1% indicated their physicians did not ask how the drug "helped" or "bothered" them after therapy was started.
- 87.7% indicated the drug was not changed after discussion of how the therapy was "helping" or "bothering" them.

a. Assume that 500 patients were interviewed in this study. Summarize each of these results using a 95% confidence interval.
b. Do you have any comments about the validity of any of these results?

Med. **10.82** People over the age of 40 years tend to notice changes in their digestive systems that alter what and how much they can eat. A study was conducted to see whether this observation applies across different ethnic segments of our society. Random samples of Anglo-Saxons, Germans, Latin Americans, Italians, Spaniards, and African Americans were obtained. The data from this survey are summarized here:

Ethnic Group	Sample Size Responding (60 of Each Group Were Contacted)	Number Reporting Altered Digestive System
Anglo-Saxon	55	7
German	58	6
Latin American	52	34
Italian	54	38
Spanish	30	20
African American	49	31

a. Does it appear that there may be a bias due to the response rates?
b. Compare the rates (π_is) for the Anglo-Saxon and German groups using a 95% confidence interval.

10.83 Refer to Exercise 10.82. There seem to be two distinct rates—those around 12% and those around 70%. Combine the sample data for the first two groups and for the last four groups. Use these data to test the hypotheses H_0: $\pi_1 - \pi_2 \geq 0$ versus H_a: $\pi_1 - \pi_2 < 0$. Here, π_1 corresponds to the population rate for the first combined group, and π_2 is the corresponding proportion for the second combined group. Give the p-value for your test.

Bus. **10.84** The following data give the observed frequencies of errors per page of unread page proofs for a sample of 40 pages from a certain journal publisher.

Errors/Page	Observed Frequencies
0	5
1	9
2	5
3	7
4	4
5	2
6	3
7	2
8	1
9	0
10	2

Conduct a test to determine whether the errors per page follow a Poisson distribution with a mean rate of 3.2. Use $\alpha = .10$.

Hort. **10.85** An entomologist was interested in studying the infestion of adult European red mites on apple trees in a Michigan orchard. She randomly selected 50 leaves from each of 10 similar apple trees in the orchard, examined the leaves, and recorded the number of mites on each of the 500 leaves. As a part of a larger study, she wanted to simulate the distribution of mites on the trees in

the orchard. Thus, the Poisson distribution was suggested as a possible model. Based on the data given here, does the Poisson distribution appear to be a plausible model for the concentration of European red mites on apple trees?

Mites per leaf	0	1	2	3	4	5	6	7
Frequency	233	127	57	33	30	10	7	3

10.86 A sample of 1,200 individuals arrested for driving under the influence of alcohol was obtained from police records. The research recorded the gender, socioeconomic status (from occupation information), and number of previous alcohol-related arrests. These data are shown here:

Socioeconomic Status	Number of Previous Alcohol-Related Arrests	Male	Female
	0	110	130
Low	1 or more	90	70
	0	105	101
Medium	1 or more	95	99
	0	90	80
High	1 or more	110	120

Separately for each socioeconomic status group, answer the following questions.

- **a.** Is there significant evidence of a difference between males and females with respect to the number of previous alcohol-related arrests?
- **b.** Compute the odds of having a previous alcohol-related arrest for both males and females. Interpret these values.
- **c.** Compute the odds ratio of having a previous alcohol-related arrest for males vesus females, and place a 95% confidence interval on the odds ratio. Interpret the interval.
- **d.** Compare the results for the three socioeconomic statuses.

10.87 Run the Mantel–Haenszel test for the above data and interpret your results.

10.88 A study was conducted to determine the relationship between annual income and number of children per family. Compute percentages for each of the income categories; then run a chi-square test of independence and draw conclusions. Use $\alpha = .10$.

Region	Number of Children per Family	Annual Income <$20,000	Annual Income ≥$20,000
East	≤ 2 children	38	67
	>2 children	220	125
South	≤ 2 children	25	78
	>2 children	120	77
West	≤ 2 children	36	66
	>2 children	95	103

Separately for each region, answer the following questions.

- **a.** Is there significant evidence of an association between annual income and number of children?
- **b.** Compute the odds ratio of having more than two children for low-income versus high-income families, and place a 95% confidence interval on the odds ratio.

c. Interpret the odds ratio.
d. Compare your results in parts (a)–(c) for the three regions.

10.89 Run the Mantel–Haenszel test for the previous data, and interpret your results.

10.90 Faculty members at a number of universities were classified according to their political ideology (left or right) and according to their academic tolerance (low, medium, or high).

Political Ideology	Academic Tolerance		
	Low	Medium	High
Left	36	44	84
Right	95	64	42

a. Is there significant evidence of an association between political ideology and academic tolerance?
b. Display the data as a graph.
c. Describe the relation between political ideology and academic tolerance.

CHAPTER 11

Linear Regression and Correlation

11.1 Introduction and Abstract of Research Study

The modeling of the relationship between a response variable and a set of explanatory variables is one of the most widely used of all statistical techniques. We refer to this type of modeling as regression analysis. A regression model provides the user with a functional relationship between the response variable and explanatory variables that allows the user to determine which of the explanatory variables have an effect on the response. The regression model allows the user to explore what happens to the response variable for specified changes in the explanatory variables. For example, financial officers must predict future cash flows based on specified values of interest rates, raw material costs, salary increases, and so on. When designing new training programs for employees, a company would want to study the relationship between employee efficiency and explanatory variables such as the results from employment tests, experience on similar jobs, educational background, and previous training. Medical researchers attempt to determine the factors that have an effect on cardiorespiratory fitness. Forest scientists study the relationship between the volume of wood in a tree and the tree's diameter at a specified height and its taper.

The basic idea of regression analysis is to obtain a model for the functional relationship between a **response variable** (often referred to as the dependent variable) and one or more **explanatory variables** (often referred to as the independent variables). Regression models have a number of uses.

1. The model provides a description of the major features of the data set. In some cases, a subset of the explanatory variables will not

affect the response variable, and, hence, the researcher will not have to measure or control any of these variables in future studies. This may result in significant savings in future studies or experiments.

2. The equation relating the response variable to the explanatory variables produced from the regression analysis provides estimates of the response variable for values of the explanatory variables not observed in the study. For example, a clinical trial is designed to study the response of a subject to various dose levels of a new drug. Because of time and budgetary constraints, only a limited number of dose levels are used in the study. The regression equation will provide estimates of the subjects' response for dose levels not included in the study. The accuracy of these estimates will depend heavily on how well the final model fits the observed data.
3. In business applications, the prediction of future sales of a product is crucial to production planning. If the data provide a model that has a good fit in relating current sales to sales in previous months, prediction of sales in future months is possible. However, a crucial element in the accuracy of these predictions is that the business conditions during which model-building data were collected remain fairly stable over the months for which the predictions are desired.
4. In some applications of regression analysis, the researcher is seeking a model that can accurately estimate the values of a variable that is difficult or expensive to measure using explanatory variables that are inexpensive to measure and obtain. If such a model is obtained, then in future applications it is possible to avoid having to obtain the values of the expensive variable by measuring the values of the inexpensive variables and using the regression equation to estimate the values of the expensive variable. For example, a physical fitness center wants to determine the physical well-being of its new clients. Maximal oxygen uptake is recognized as the single best measure of cardiorespiratory fitness, but its measurement is expensive. Therefore, the director of the fitness center would want a model that provides accurate estimates of maximal oxygen uptake using easily measured variables such as weight, age, heart rate after a 1-mile walk, time needed to walk 1 mile, and so on.

prediction versus explanation

We can distinguish between **prediction** (reference to future values) and **explanation** (reference to current or past values). Because of the virtues of hindsight, explanation is easier than prediction. However, it is often clearer to use the term *prediction* to include both cases. Therefore, in this book, we sometimes blur the distinction between prediction and explanation.

For prediction (or explanation) to make much sense, there must be some connection between the variable we're predicting (the dependent variable) and the variable we're using to make the prediction (the independent variable). No doubt, if you tried long enough, you could find 30 common stocks whose price changes over a year have been accurately predicted by the won–lost percentage of the 30 major league baseball teams on the fourth of July. However, such a prediction is absurd because there is no connection between the two variables. Prediction requires a

unit of association

unit of association; there should be an entity that relates the two variables. With time-series data, the unit of association may simply be time. The variables may be measured at the same time period, or for genuine prediction, the independent

variable may be measured at a time period before the dependent variable. For cross-sectional data, an economic or physical entity should connect the variables. If we are trying to predict the change in market share of various soft drinks, we should consider the promotional activity for those drinks, not the advertising for various brands of spaghetti sauce. The need for a unit of association seems obvious, but many predictions are made for situations in which no such unit is evident.

simple linear regression

In this chapter, we consider **simple linear regression** analysis, in which there is a single given independent variable x and the equation for predicting a dependent variable y is a linear function of that independent variable. Suppose, for example, that the director of a county highway department wants to predict the cost of a resurfacing contract that is up for bids. We could reasonably predict the costs to be a function of the road miles to be resurfaced. A reasonable first attempt is to use a linear production function. Let y = total cost of a project in thousands of dollars, x = number of miles to be resurfaced, and $\hat{y}$ = total predicted cost, also in thousands of dollars. The prediction equation $\hat{y} = 2.0 + 3.0x$ (for example) is a linear equation. The constant term, such as the 2.0, is the **intercept** term and is interpreted as the predicted value of y when $x = 0$. In the road-resurfacing example, we may interpret the intercept as the fixed cost of beginning the project. The coefficient of x, such as the 3.0, is the **slope** of the line, the predicted change in y when there is a one-unit change in x. In the road-resurfacing example, if two projects differed by 1 mile in length, we would predict that the longer project would cost 3 (thousand dollars) more than the shorter one. In general, we write the prediction equation as

intercept

slope

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

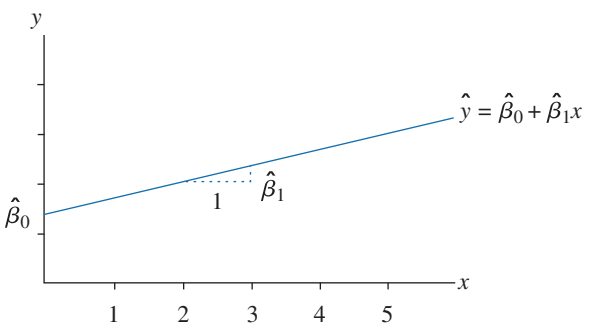
where $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1$ is the slope. See Figure 11.1.

The basic idea of simple linear regression is to use data to fit a prediction line that relates a dependent variable y and a single independent variable x. The first assumption in simple regression is that the relation is in fact linear. According to the **assumption of linearity,** the slope of the equation does not change as x changes. In the road-resurfacing example, we assume that there would be no (substantial) economies or diseconomies from projects of longer mileage. There is little point in using simple linear regression unless the linearity assumption makes sense (at least roughly).

assumption of linearity

Linearity is not always a reasonable assumption on its face. For example, if we tried to predict y = number of drivers that are aware of a car dealer's midsummer sale using x = number of repetitions of the dealer's radio commercial, the assumption of linearity means that the first broadcast of the commercial leads to no greater an increase in aware drivers than the thousand-and-first broadcast.

FIGURE 11.1
Linear prediction function



(You've heard commercials like that.) We strongly doubt that such an assumption is valid over a wide range of x-values. It makes far more sense to us that the effect of repetition would diminish as the number of repetitions got larger, so a straight-line prediction wouldn't work well.

Assuming linearity, we would like to write y as a linear function of x: $y = \beta_0 + \beta_1 x$. However, according to such an equation, y is an exact linear function of x; no room is left for the inevitable errors (deviation of actual y-values from their predicted values). Therefore, corresponding to each y, we introduce a **random error term** ε_i and assume the model

random error term

$$y = \beta_0 + \beta_1 x + \varepsilon$$

We assume the random variable y to be made up of a predictable part (a linear function of x) and an unpredictable part (the random error ε_i). The coefficients β_0 and β_1 are interpreted as the true, underlying intercept and slope. The error term ε includes the effects of all other factors, known or unknown. In the road-resurfacing project, unpredictable factors such as strikes, weather conditions, and equipment breakdowns would contribute to ε, as would factors such as hilliness or prerepair condition of the road—factors that might have been used in prediction but were not. The combined effects of unpredictable and ignored factors yield the random error terms ε.

For example, one way to predict the gas mileage of various new cars (the dependent variable) based on their curb weight (the independent variable) would be to assign each car to a different driver, say, for a 1-month period. What unpredictable and ignored factors might contribute to prediction error? Unpredictable (random) factors in this study would include the driving habits and skills of the drivers, the type of driving done (city versus highway), and the number of stoplights encountered. Factors that would be ignored in a regression analysis of mileage and weight would include engine size and type of transmission (manual versus automatic).

In regression studies, the values of the independent variable (the x_i values) are usually taken as predetermined constants, so the only source of randomness is the ε_i terms. Although most economic and business applications have fixed x_i values, this is not always the case. For example, suppose that x_i is the score of an applicant on an aptitude test and y_i is the productivity of the applicant. If the data are based on a random sample of applicants, x_i (as well as y_i) is a random variable. The question of fixed versus random in regard to x is not crucial for regression studies. If the x_is are random, we can simply regard all probability statements as conditional on the observed x_is.

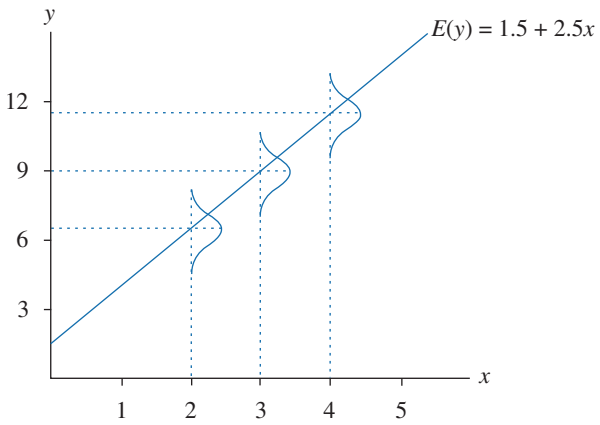
When we assume that the x_is are constants, the only random portion of the model for y_i is the random error term ε_i. We make the following formal assumptions.

DEFINITION 11.1

Formal assumptions of regression analysis:

1. The relation is in fact linear, so that the errors all have expected values of zero: $E(\varepsilon_i) = 0$ for all i.
2. The errors all have the same variance: $\text{Var}(\varepsilon_i) = \sigma^2$ for all i.
3. The errors are independent of each other.
4. The errors are all normally distributed; ε_i is normally distributed for all i.

FIGURE 11.2
Theoretical distribution of y in regression



These assumptions are illustrated in Figure 11.2. The actual values of the dependent variable are distributed normally with mean values falling on the regression line and the same standard deviation at all values of the independent variable. The only assumption not shown in the figure is independence from one measurement to another.

These are the formal assumptions, made in order to derive the significance tests and prediction methods that follow. We can begin to check these assumptions

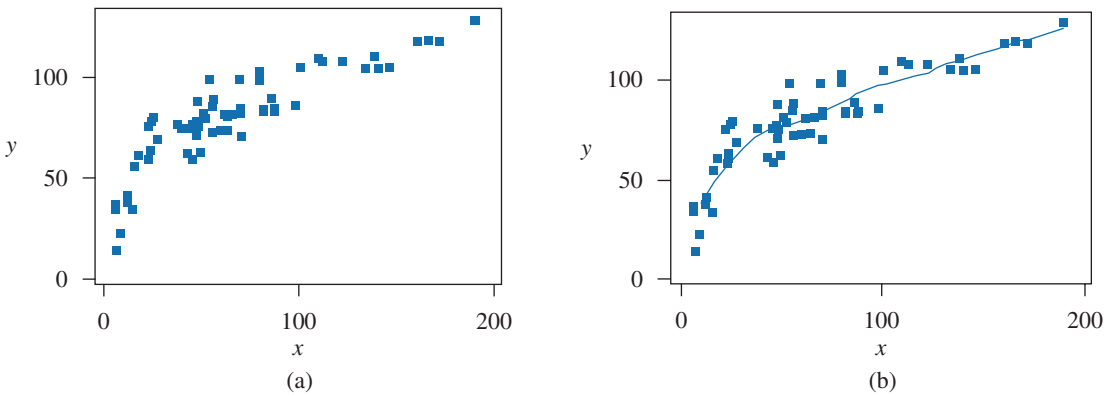
scatterplot

by looking at a **scatterplot** of the data. This is simply a plot of each (x, y) point, with the independent variable value measured on the horizontal axis and the dependent variable value measured on the vertical axis. Look to see whether the points basically fall around a straight line or whether there is a definite curve in the pattern. Also look to see whether there are any evident outliers falling far from the general pattern of the data. A scatterplot is shown in part (a) of Figure 11.3.

smoothers

There are a number of nonparametric **smoothers,** which will sketch a curve through data without necessarily assuming any particular model. If such a smoother yields something close to a straight line, then linear regression is reasonable. One such method is called LOWESS (locally weighted scatterplot smoother). Roughly, a smoother takes a relatively narrow "slice" of data along the x axis, calculates

FIGURE 11.3 (a) Scatterplot. (b) LOWESS curve.



a line that fits the data in that slice, moves the slice slightly along the x axis, recalculates the line, and so on. Then all the little lines are connected in a smooth curve. The width of the slice is called the *bandwidth;* this may often be controlled in the computer program that does the smoothing. The plain scatterplot (Figure 11.3a) is shown again (Figure 11.3b) with a LOWESS curve through it. The scatterplot shows a curved relation; the LOWESS curve confirms that impression.

spline fit

Another type of scatterplot smoother is the **spline fit.** It can be understood as taking a narrow slice of data, fitting a curve (often a cubic equation) to the slice, moving to the next slice, fitting another curve, and so on. The curves are calculated in such a way as to form a connected, continuous curve.

transformation

Many economic relations are not linear. For example, any diminishing returns pattern will tend to yield a relation that increases—but at a decreasing rate. If the scatterplot does not appear linear, by itself or when fitted with a LOWESS curve, it can often be "straightened out" by a **transformation** of either the independent variable or the dependent variable. A good statistical computer package or a spreadsheet program will compute such functions as the square root of each value of a variable. The transformed variable should be thought of as simply another variable.

For example, a large city dispatches crews each spring to patch potholes in its streets. Records are kept of the number of crews dispatched each day and the number of potholes filled that day. A scatterplot of the number of potholes patched and the number of crews dispatched and the same scatterplot with a LOWESS curve through it are shown in Figure 11.4. The relation is not linear. Even without the LOWESS curve, the decreasing slope is obvious. That's not surprising; as the city sends out more crews, they will be using less effective workers, the crews will have to travel farther to find holes, and so on. All these reasons suggest that diminishing returns will occur.

We can try several transformations of the independent variable to find a scatterplot in which the points more nearly fall along a straight line. Three common transformations are square root, natural logarithm, and inverse (1 divided by the variable). We applied each of these transformations to the pothole repair data. The results are shown in Figures 11.5a–c with LOWESS curves. The square root transformation (a) and inverse transformation (c) didn't really give us a

FIGURE 11.4 Scatterplots for pothole data

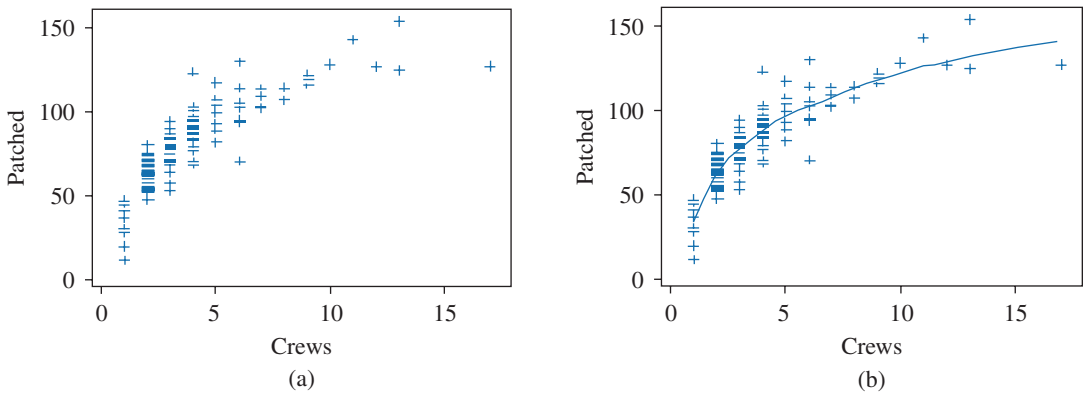
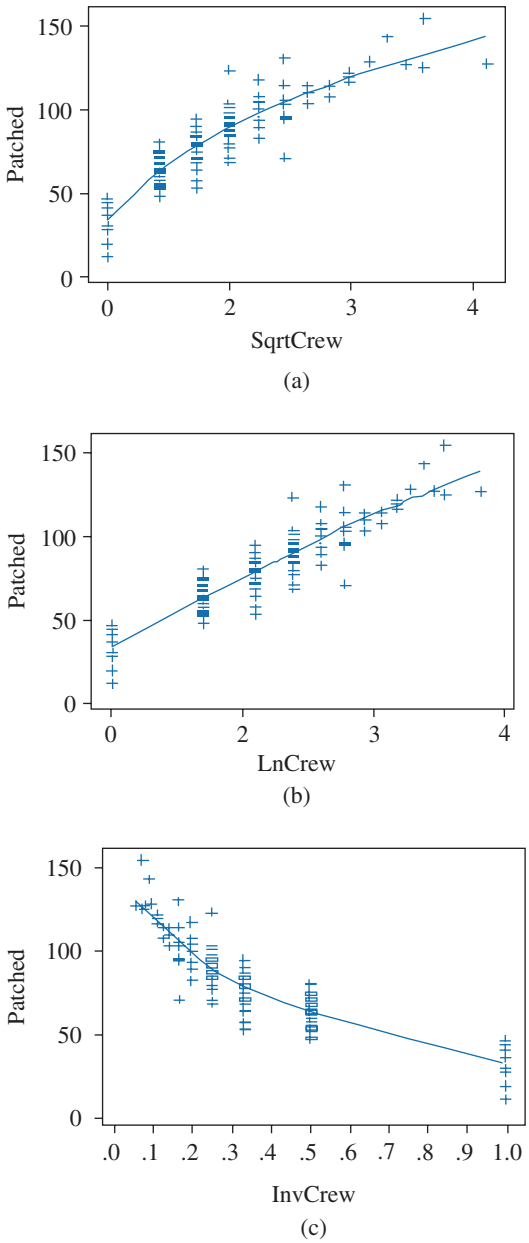


FIGURE 11.5
Scatterplots with transformed predictor



straight line. The natural logarithm (b) worked very well, however. Therefore, we would use LnCrew as our independent variable.

Finding a good transformation often requires trial and error. Following are some suggestions to try for transformations. Note that there are *two* key features to look for in a scatterplot. First, is the relation nonlinear? Second, is there a pattern of increasing variability along the *y* (vertical) axis? If there is, the assumption of constant variance is questionable. These suggestions don't cover all the possibilities but do include the most common problems.

DEFINITION 11.2

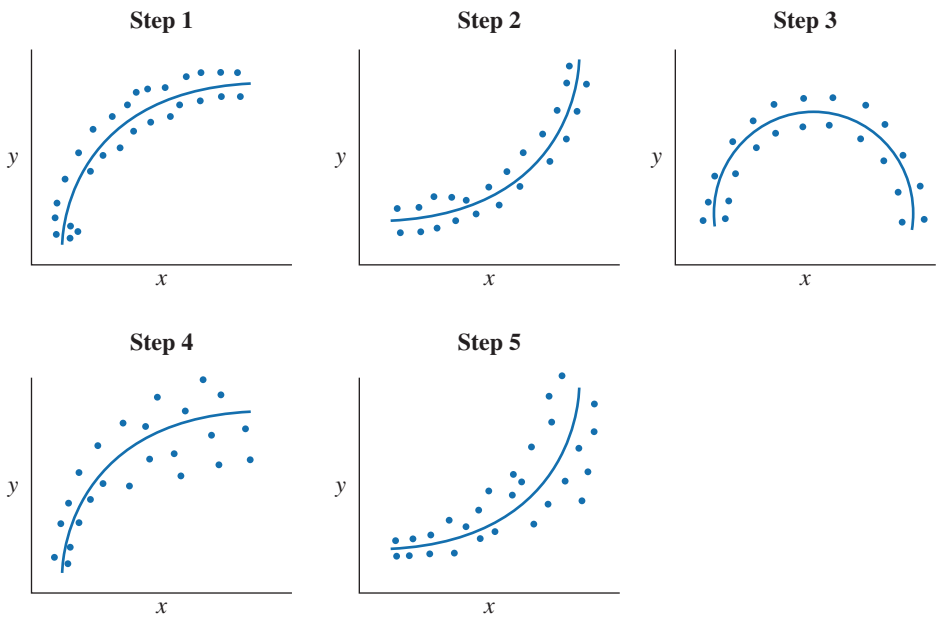
Steps for choosing a transformation:

1. If the plot indicates a relation that is increasing but at a decreasing rate and if variability around the curve is roughly constant, transform x using a square root, logarithm, or inverse transformation.
2. If the plot indicates a relation that is increasing at an increasing rate and if variability is roughly constant, try using both x and x^2 as predictors. Because this method uses two variables, the multiple regression methods of the next two chapters are needed.
3. If the plot indicates a relation that increases to a maximum and then decreases and if variability around the curve is roughly constant, again try using both x and x^2 as predictors.
4. If the plot indicates a relation that is increasing at a decreasing rate and if variability around the curve increases as the predicted y-value increases, try using y^2 as the dependent variable.
5. If the plot indicates a relation that is increasing at an increasing rate and if variability around the curve increases as the predicted y-value increases, try using $\ln(y)$ as the dependent variable. It sometimes may also be helpful to use $\ln(x)$ as the independent variable. Note that a change in a natural logarithm corresponds quite closely to a percentage change in the original variable. Thus, the slope of a transformed variable can be interpreted quite well as a percentage change.

The plots in Figure 11.6 correspond to the descriptions given in Definition 11.2.

There are symmetric recommendations for the situations where the relation is decreasing at a decreasing rate (use Step 1 or Step 4 transformations) or where the relation is decreasing at an increasing rate (use Step 2 or Step 5 transformations).

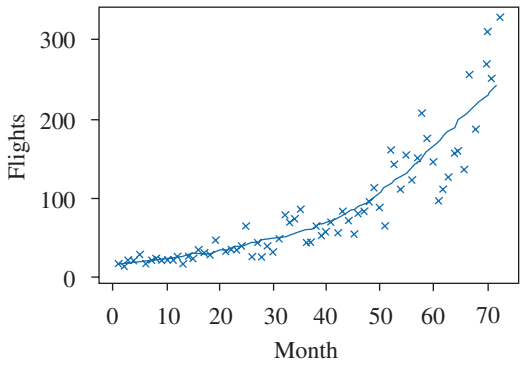
FIGURE 11.6
Plots corresponding to steps in Definition 11.2



EXAMPLE 11.1

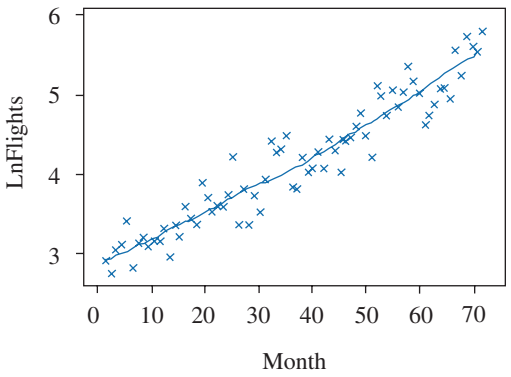
An airline has seen a very large increase in the number of free flights used by participants in its frequent flyer program. To try to predict the trend in these flights in the near future, the director of the program assembled data for the last 72 months. The dependent variable y is the number of thousands of free flights; the independent variable x is the month number. A scatterplot with a LOWESS smoother, done using Minitab, is shown in Figure 11.7. What transformation is suggested?

FIGURE 11.7
Frequent flyer free flights by month



Solution The pattern shows flights increasing at an increasing rate. The LOWESS curve is definitely turning upward. In addition, variation (up and down) around the curve is increasing. The points around the high end of the curve (on the right, in this case) scatter much more than the ones around the low end of the curve. The increasing variability suggests transforming the y-variable. A natural logarithm (ln) transformation often works well. Minitab computed the logarithms and replotted the data, as shown in Figure 11.8. The pattern is much closer to a straight line, and the scatter around the line is much closer to constant.

FIGURE 11.8
Result of logarithm transformation



We will have more to say about checking assumptions in Chapter 12. For a simple regression with a single predictor, careful checking of a scatterplot, ideally with a smooth curve fit through it, will help avoid serious blunders.

Once we have decided on any mathematical transformations, we must estimate the actual equation of the regression line. In practice, only sample data are available. The population intercept, slope, and error variance all have to be estimated from limited sample data. The assumptions we made in this section allow us to make inferences about the true parameter values from the sample data. ■

Abstract of Research Study: Two Methods for Detecting *E. coli*

The case study in Chapter 7 described a new microbial method for the detection of *E. coli,* the Petrifilm HEC test. The researchers wanted to evaluate the agreement of the results obtained using the HEC test with results obtained from an elaborate laboratory-based procedure, hydrophobic grid membrane filtration (HGMF). The HEC test is easier to inoculate, more compact to incubate, and safer to handle than conventional procedures. However, prior to using the HEC procedure, it was necessary to compare the readings from the HEC test to the readings from the HGMF procedure obtained on the same meat sample to determine whether the two procedures were yielding the same readings. If the readings differed but an equation could be obtained that could closely relate the HEC reading to the HGMF reading, then the researchers could calibrate the HEC readings to predict what readings would have been obtained using the HGMF test procedure. If the HEC test results were unrelated to the HGMF test procedure results, then the HEC test could not be used in the field in detecting *E. coli*. The necessary regression analysis to answer these questions will be given at the end of this chapter.

11.2 Estimating Model Parameters

The intercept β_0 and slope β_1 in the regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

are population quantities. We must estimate these values from sample data. The error variance σ_ε^2 is another population parameter that must be estimated. The first regression problem is to obtain estimates of the slope, intercept, and variance; we discuss how to do so in this section.

The road-resurfacing example of Section 11.1 is a convenient illustration. Suppose the following data for similar resurfacing projects in the recent past are available. Note that we do have a unit of association: The connection between a particular cost and mileage is that they're based on the same project.

Cost y_i (in thousands of dollars):	6.0	14.0	10.0	14.0	26.0
Mileage x_i (in miles):	1.0	3.0	4.0	5.0	7.0

A first step in examining the relation between y and x is to plot the data as a scatterplot. Remember that each point in such a plot represents the (x, y) coordinates of one data entry, as in Figure 11.9. The plot makes it clear that there is an imperfect but generally increasing relation between x and y. A straight-line relation appears plausible; there is no evident transformation with such limited data.

The regression analysis problem is to find the best straight-line prediction. The most common criterion for "best" is based on squared prediction error. We find the equation of the prediction line—that is, the slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ that minimize the total squared prediction error. The method that accomplishes this goal is called the **least-squares method** because it chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the quantity:

least-squares method

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

FIGURE 11.9
Scatterplot of cost versus mileage

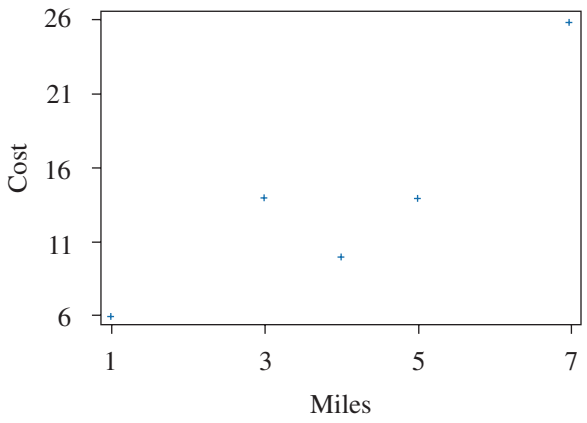
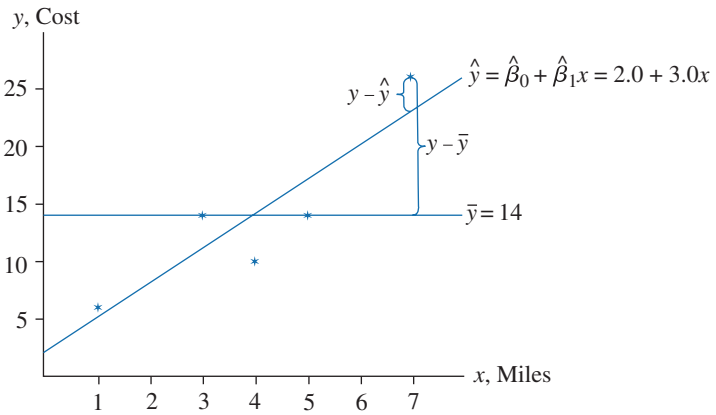


FIGURE 11.10
Deviations from the least-squares line from the mean



The prediction errors are shown on the plot of Figure 11.10 as vertical deviations from the line. The deviations are taken as vertical distances because we're trying to predict y-values and errors should be taken in the y direction. For these data, the least-squares line can be shown to be $\hat{y} = 2.0 + 3.0x$; one of the deviations from it is indicated by the smaller brace. For comparison, the mean $\bar{y} = 14.0$ is also shown; deviation from the mean is indicated by the larger brace. The least-squares principle leads to some fairly long computations for the slope and intercept. Usually, these computations are done by computer.

DEFINITION 11.3

The **least-squares estimates of slope and intercept** are obtained as follows:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \quad \text{and} \quad S_{xx} = \sum_i (x_i - \bar{x})^2$$

Thus, S_{xy} is the sum of x deviations times y deviations, and S_{xx} is the sum of x deviations squared.

For the road-resurfacing data, $n = 5$ and

$$\sum x_i = 1.0 + \ldots + 7.0 = 20.0$$

so $\bar{x} = \frac{20.0}{5} = 4.0$. Similarly,

$$\sum y_i = 70.0$$

so $\bar{y} = \frac{70.0}{5} = 14.0$.

Also,

$$\begin{aligned} S_{xx} &= \sum (x_i - \bar{x})^2 \\ &= (1.0 - 4.0)^2 + \cdots + (7.0 - 4.0)^2 \\ &= 20.00 \end{aligned}$$

and

$$\begin{aligned} S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) \\ &= (1.0 - 4.0)(6.0 - 14.0) + \cdots + (7.0 - 4.0)(26.0 - 14.0) \\ &= 60.0 \end{aligned}$$

Thus,

$$\hat{\beta}_1 = \frac{60.0}{20.0} = 3.0 \text{ and } \hat{\beta}_0 = 14.0 - (3.0)(4.0) = 2.0$$

From the value $\hat{\beta}_1 = 3$, we can conclude that the estimated average increase in cost for each additional mile is \$3,000.

EXAMPLE 11.2

Data from a sample of 10 pharmacies are used to examine the relation between prescription sales volume and percentage of prescription ingredients purchased directly from the supplier. The sample data are shown in Table 11.1.

TABLE 11.1
Data for Example 11.2

Pharmacy	Sales Volume, y (in \$1,000s)	% of Ingredients Purchased Directly, x
1	25	10
2	55	18
3	50	25
4	75	40
5	110	50
6	138	63
7	90	42
8	60	30
9	10	5
10	100	55

a. Find the least-squares estimates for the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
b. Predict sales volume for a pharmacy that purchases 15% of its prescription ingredients directly from the supplier.
c. Plot the (x, y) data and the prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
d. Interpret the value of $\hat{\beta}_1$ in the context of the problem.

Solution

a. The equation can be calculated by virtually any statistical computer package; for example, here is abbreviated Minitab output:

```
Regression Analysis: Sales versus PurchDirect

Analysis of Variance

Source          DF   Adj SS   Adj MS  F-Value  P-Value
Regression       1  13231.0  13231.0   162.56    0.000
Error            8    651.1     81.4
Total            9  13882.1

Model Summary

      S     R-sq   R-sq(adj)   R-sq(pred)
9.02171   95.31%      94.72%       92.32%

Coefficients

Term          Coef   SE Coef  T-Value  P-Value      VIF
Constant      4.70      5.95     0.79    0.453
PurchDirect 1.970     0.155    12.75    0.000     1.00

Regression Equation:  Sales = 4.70 + 1.970 PurchDirect
```

To see how the computer does the calculations, you can obtain the least-squares estimates from Table 11.2.

TABLE 11.2
Calculations for obtaining least-squares estimates

	y	x	$y - \bar{y}$	$x - \bar{x}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
	25	10	−46.3	−23.8	1,101.94	566.44
	55	18	−16.3	−15.8	257.54	249.64
	50	25	−21.3	−8.8	187.44	77.44
	75	40	3.7	6.2	22.94	38.44
	110	50	38.7	16.2	626.94	262.44
	138	63	66.7	29.2	1,947.64	852.64
	90	42	18.7	8.2	153.34	67.24
	60	30	−11.3	−3.8	42.94	14.44
	10	5	−61.3	−28.8	1,765.44	829.44
	100	55	28.7	21.2	608.44	449.44
Total	713	338	0	0	6,714.60	3,407.60
Mean	71.3	33.8				

$$S_{xx} = \sum (x - \bar{x})^2 = 3{,}407.6$$

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = 6{,}714.6$$

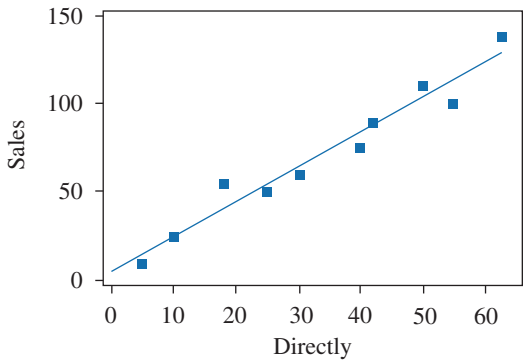
Substituting into the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{6{,}714.6}{3{,}407.6} = 1.9704778 \quad \text{rounded to } 1.97$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 71.3 - 1.9704778(33.8) = 4.6978519 \quad \text{rounded to } 4.70$$

b. When $x = 15\%$, the predicted sales volume is $\hat{y} = 4.70 + 1.97(15) = 34.25$ (that is, \$34,250).
c. The (x, y) data and prediction line are plotted in Figure 11.11.
d. From $\hat{\beta}_1 = 1.97$, we conclude that if a pharmacy would increase by 1% the percentage of ingredients purchased directly, then the estimated increase in average sales volume would be \$1,970.

FIGURE 11.11
Sample data and least-squares prediction line



EXAMPLE 11.3

In Chapter 3, we discussed a study that related the crime rate in a major city to the number of casino employees in that city. The study was attempting to associate an increasing crime rate with increasing levels of casino gambling, which is reflected in the number of people employed in the gambling industry. Use the information in Table 3.18 to calculate the least-squares estimates of the intercept and slope of the line relating crime rate to number of casino employees. Use the Minitab output below to confirm your calculations.

Solution From Table 3.18, we have the following summary statistics for the crime rate y (number of crimes per 1,000 population) and the number of casino employees x (in thousands):

$$\bar{x} = \frac{318}{10} = 31.80, \quad \bar{y} = \frac{27.85}{10} = 2.785$$

$$S_{xx} = 485.60, \quad S_{yy} = 7.3641, \quad S_{xy} = 55.810$$

Thus,

$$\hat{\beta}_1 = \frac{55.810}{485.60} = .11493 \quad \text{and} \quad \hat{\beta}_0 = 2.785 - (.11493)(31.80) = -.8698$$

The Minitab output is given here

```
Regression Analysis: CrimeRate versus NumEmployees

Analysis of Variance

Source          DF  Adj SS  Adj MS  F-Value  P-Value
Regression       1  6.4142  6.4142    54.03    0.000
Error            8  0.9498  0.1187
Total            9  7.3641

Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.344566  87.10%     85.49%      81.84%

Coefficients

Term            Coef  SE Coef  T-Value  P-Value   VIF
Constant      -0.870    0.509    -1.71    0.126
NumEmployees  0.1149   0.0156     7.35    0.000  1.00

Regression Equation:  CrimeRate = -0.870 + 0.1149 NumEmployees
```

From the previous output, the values calculated are the same as the values from Minitab. We would interpret the value of the estimated slope $\hat{\beta}_1 = .1149$ as follows. For an increase of 1,000 employees in the casino industry, the average crime rate would increase .115. It is important to note that these types of social relationships are much more complex than this simple relationship. Also, it would be a major mistake to place much credence in this type of conclusion because of all the other factors that may have an effect on the crime rate. ■

high leverage point

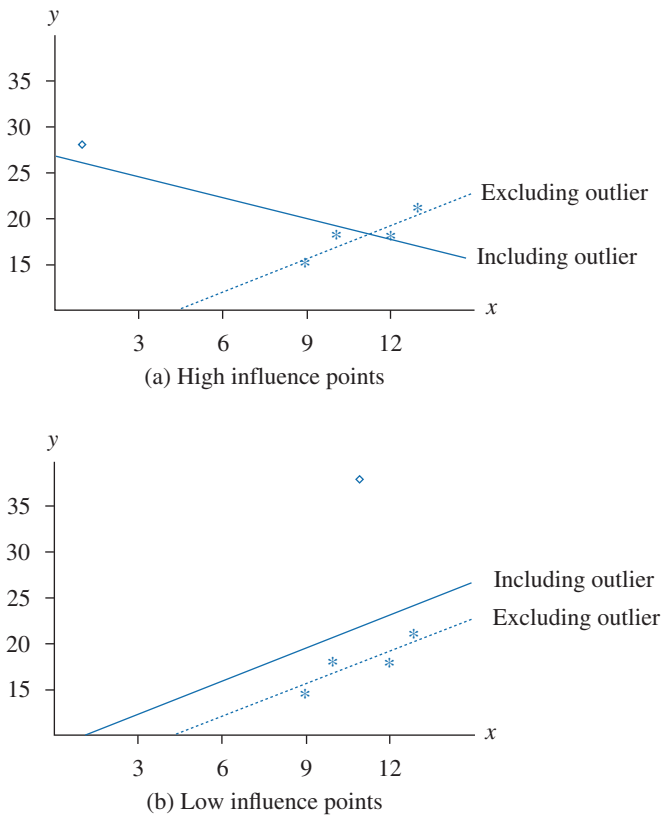
high influence point

The estimate of the regression slope can potentially be greatly affected by **high leverage points.** These are points that have very high or very low values of the independent variable—outliers in the x direction. They carry great weight in the estimate of the slope. A high leverage point that also happens to correspond to a y outlier is a **high influence point.** It will alter the slope and twist the line badly.

A point has high influence if omitting it from the data will cause the regression line to change substantially. To have high influence, a point must first have high leverage and must, in addition, fall outside the pattern of the remaining points. Consider the two scatterplots in Figure 11.12. In plot (a), the point in the upper left corner is far to the left of the other points; it has a much lower x-value and therefore has high leverage. If we drew a line through the other points, the line would fall far below this point, so the point is an outlier in the y direction as well. Therefore, it also has high influence. Including this point would change the slope of the line greatly. In contrast, in plot (b), the y outlier point corresponds to an x-value very near the mean, having low leverage. Including this point would pull the line upward, increasing the intercept, but it wouldn't increase or decrease the slope much at all. Therefore, it does not have great influence.

A high leverage point indicates only a *potential* distortion of the equation. Whether or not including the point will "twist" the equation depends on its influence (whether or not the point falls near the line through the remaining points). A point must have *both* high leverage and an outlying y-value to qualify as a high influence point.

FIGURE 11.12
High versus low influence



Mathematically, the effect of a point's leverage can be seen in the S_{xy} term, which enters into the slope calculation. One of the many ways this term can be written is

$$S_{xy} = \sum (x_i - \bar{x}) y_i$$

We can think of this equation as a weighted sum of y-values. The weights are large positive or negative numbers when the x-value is far from its mean and has high leverage. The weight is almost 0 when x is very close to its mean and has low leverage.

diagnostic measures

Most computer programs that perform regression analyses will calculate one or another of several **diagnostic measures** of leverage and influence. We won't try to summarize all of these measures. We only note that very large values of any of these measures correspond to very high leverage or influence points. The distinction between high leverage (x outlier) and high influence (x outlier and y outlier) points is not universally agreed upon yet. Check the program's documentation to see what definition is being used.

The standard error of the slope $\hat{\beta}_1$ is calculated by all statistical packages. Typically, it is shown in output in a column to the right of the coefficient column. Like any standard error, it indicates how accurately one can estimate the correct population or process value. The quality of estimation of $\hat{\beta}_1$ is influenced by two

quantities: the error variance σ_ε^2 and the amount of variation in the independent variable S_{xx}:

$$\sigma_{\hat{\beta}_1} = \frac{\sigma_\varepsilon}{\sqrt{S_{xx}}}$$

The greater the variability σ_ε of the y-value for a given value of x, the larger the value of $\sigma_{\hat{\beta}_1}$. Sensibly, if there is high variability around the regression line, it is difficult to estimate that line. Also, the smaller the variation in x-values (as measured by S_{xx}), the larger the value of $\sigma_{\hat{\beta}_1}$. The slope is the predicted change in y per unit change in x; if x changes very little in the data, so that S_{xx} is small, it is difficult to estimate the rate of change in y accurately. If the price of a brand of diet soda has not changed for years, it is obviously hard to estimate the change in quantity demanded when price changes. The standard error of the estimated intercept $\hat{\beta}_0$ is influenced by n, naturally, and also by the size of the square of the sample mean, $\bar{x}^2$, relative to S_{xx}.

$$\sigma_{\hat{\beta}_0} = \sigma_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

The intercept is the predicted y-value when $x = 0$; if all the x_i are, for instance, large positive numbers, predicting y at $x = 0$ is a huge extrapolation from the actual data. Such extrapolation magnifies small errors, and the standard error of $\hat{\beta}_0$ is large. The ideal situation for estimating $\hat{\beta}_0$ is when $\bar{x} = 0$.

To this point, we have considered only the estimates of intercept and slope. We also have to estimate the true error variance σ_ε^2. We can think of this quantity as "variance around the line" or as the mean squared prediction error. The estimate of σ_ε^2 is based on the **residuals** $y_i - \hat{y}_i$, which are the prediction errors in the sample. The estimate of σ_ε^2 based on the sample data is the sum of squared residuals divided by $n - 2$, the degrees of freedom. The estimated variance is often shown in computer output as MSE(Error) or MS(Residual). Recall that MS stands for "mean square" and is always a sum of squares divided by the appropriate degrees of freedom:

residuals

$$s_\varepsilon^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n - 2} = \frac{\text{SS(Error)}}{n - 2}$$

In the computer output for Example 11.3, SS(Error) is shown to be 0.9498.

Just as we divide by $n - 1$ rather than by n in the ordinary sample variance s^2 (in Chapter 3), we divide by $n - 2$ in s_ε^2, the estimated variance around the line. The reduction from n to $n - 2$ occurs because in order to estimate the variability around the regression line, we must first estimate the two parameters, β_0 and β_1, to obtain the estimated line. The effective sample size for estimating σ_ε^2 is thus $n - 2$. In our definition, s_ε^2 is undefined for $n = 2$, as it should be. Another argument is that dividing by $n - 2$ makes s_ε^2 an unbiased estimator of σ_ε^2. In the computer output of Example 11.3, $n - 2 = 10 - 2 = 8$ is shown as DF (degrees of freedom) for Error and $s_\varepsilon^2 = 0.1187$ is shown as MS for Error.

The square root s_ε of the sample variance is called the sample standard deviation around the regression line, the standard error of estimate, or the **residual standard deviation.** Because s_ε estimates σ_ε, the standard deviation of y_i, σ_ε estimates the standard deviation of the population of y-values associated with a given value of the independent variable x. The output in Example 11.3 labels s_ε as S with $S = 0.344566$.

residual standard deviation

Like any other standard deviation, the residual standard deviation may be interpreted by the Empirical Rule. About 95% of the prediction errors will fall within ±2 standard deviations of the mean error; the mean error is always 0 in the least-squares regression model. Therefore, a residual standard deviation of 0.345 means that about 95% of prediction errors will be less than $\pm 2(0.345) = \pm 0.690$.

The estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and s_ε are basic in regression analysis. They specify the regression line and the probable degree of error associated with y-values for a given value of x. The next step is to use these sample estimates to make inferences about the true parameters.

EXAMPLE 11.4

Forest scientists are concerned with the decline in forest growth throughout the world. One aspect of this decline is the possible effect of emissions from coal-fired power plants. The scientists in particular are interested in the pH level of the soil and the resulting impact on tree growth retardation. The scientists study various forests that are likely to be exposed to these emissions. They measure various aspects of growth associated with trees in a specified region and the soil pH in the same region. The forest scientists then want to determine impact on tree growth as the soil becomes more acidic. An index of growth retardation is constructed from the various measurements taken on the trees with a high value indicating greater retardation in tree growth. A higher value of soil pH indicates a more acidic soil. Twenty tree stands that are exposed to the power plant emissions are selected for study. The values of the growth retardation index and average soil pH are recorded in Table 11.3.

TABLE 11.3
Forest growth retardation data

Stand	Soil pH	Grow Ret	Stand	Soil pH	Grow Ret
1	3.3	17.78	11	3.9	14.95
2	3.4	21.59	12	4.0	15.87
3	3.4	23.84	13	4.1	17.45
4	3.5	15.13	14	4.2	14.35
5	3.6	23.45	15	4.3	14.64
6	3.6	20.87	16	4.4	17.25
7	3.7	17.78	17	4.5	12.57
8	3.7	20.09	18	5.0	7.15
9	3.8	17.78	19	5.1	7.50
10	3.8	12.46	20	5.2	4.34

The scientists expect that as the soil pH increases within an acceptable range, the trees will have a lower growth retardation index value.

Using the above data and Minitab, do the following:

a. Examine the scatterplot and decide whether a straight line is a reasonable model.
b. Identify least-squares estimates for β_0 and β_1 in the model $y = \beta_0 + \beta_1 x + \varepsilon$, where y is the index of growth retardation and x is the soil pH.
c. Predict the growth retardation for a soil pH of 4.0.
d. Identify s_ε, the sample standard deviation about the regression line.
e. Interpret the value of $\hat{\beta}_1$.

```
Regression Analysis: GrowthRet versus SoilPh

Analysis of Variance

Source          DF  Adj SS   Adj MS  F-Value  P-Value
Regression       1  385.28  385.276    52.01    0.000
Error           18  133.33    7.407
Total           19  518.61

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
2.72162  74.29%     72.86%      68.59%

Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  47.48     4.43    10.72    0.000
SoilPh    -7.86     1.09    -7.21    0.000  1.00

Regression Equation:   GrowthRet = 47.48 - 7.86 SoilPh
```

Solution

a. A scatterplot drawn by the Minitab package is shown in Figure 11.13. The data appear to fall approximately along a downward-sloping line. There does not appear to be a need for using a more complex model.

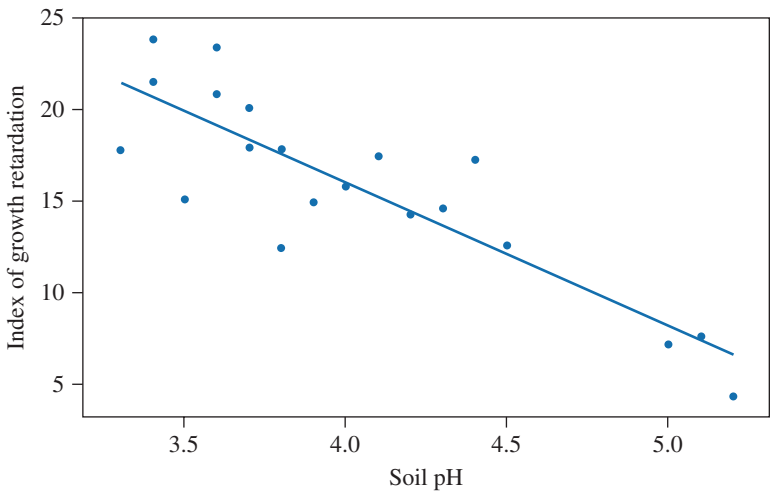


FIGURE 11.13
Scatterplot of growth retardation versus soil pH

b. The output shows the coefficients twice, with differing numbers of digits. The estimated intercept (constant) is $\hat{\beta}_0 = 47.48$, and the estimated slope (Soil pH) is $\hat{\beta}_1 = -7.86$. Note that the negative slope corresponds to a downward-sloping line.

c. The least-squares prediction when $x = 4.0$ is

$$\hat{y} = 47.48 - 7.86(4.0) = 16.04$$

d. The standard deviation around the fitted line (the residual standard deviation) is shown as $S = 2.72162$.

e. From $\hat{\beta}_1 = -7.86$, we conclude that for a one-unit increase in soil pH, there is an estimated decrease of 7.86 in the average value of the growth retardation index. ■

11.3 Inferences About Regression Parameters

The slope, intercept, and residual standard deviation in a simple regression model are all estimates based on limited data. As with all other statistical quantities, they are affected by random error. In this section, we consider how to allow for that random error. The concepts of hypothesis tests and confidence intervals that we have applied to means and proportions apply equally well to regression summary figures.

t test for β_1

The t distribution can be used to make significance tests and confidence intervals for the true slope and intercept. One natural null hypothesis is that the true slope β_1 equals 0. If this H_0 is true, a change in x yields no predicted change in y, and it follows that x has no value in predicting y. We know from the previous section that the sample slope $\hat{\beta}_1$ has the expected value β_1 and standard error

$$\sigma_{\hat{\beta}_1} = \sigma_\varepsilon \sqrt{\frac{1}{S_{xx}}}$$

In practice, σ_ε is not known and must be estimated by s_ε, the residual standard deviation. In almost all regression analysis computer outputs, the estimated standard error is shown next to the coefficient. A test of this null hypothesis is given by the t statistic

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{estimated standard error } (\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{s_\varepsilon \sqrt{1/S_{xx}}}$$

The most common use of this statistic is shown in the following summary.

Summary of a Statistical Test for β_1

Hypotheses:

Case 1. H_0: $\beta_1 \le 0$ versus H_a: $\beta_1 > 0$
Case 2. H_0: $\beta_1 \ge 0$ versus H_a: $\beta_1 < 0$
Case 3. H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \ne 0$

T.S.: $t = \dfrac{\hat{\beta}_1 - 0}{s_\varepsilon / \sqrt{S_{xx}}}$

R.R.: For df $= n - 2$ and Type I error α,

1. Reject H_0 if $t > t_\alpha$.
2. Reject H_0 if $t < -t_\alpha$.
3. Reject H_0 if $|t| > t_{\alpha/2}$.

Check assumptions and draw conclusions.

All regression analysis outputs show this t-value.

In most computer outputs, this test is indicated after the standard error and labeled as T-Value or T STATISTIC. Often, a p-value is also given, which eliminates the need for looking up the t-value in a table.

EXAMPLE 11.5

Use the computer output of Example 11.4 to locate the value of the t statistic for testing H_0: $\beta_1 = 0$ in the tree growth retardation example. Give the observed level of significance for the test.

Solution From the Minitab output, the value of the test statistic is $t = -7.21$. The p-value for the two-tailed alternative H_a: $\beta_1 \neq 0$, labeled as P, is .000. In fact, the value is given by p-value $= 2P(t > 7.21) = 2(1 - \mathbf{pt}(7.21, 18)) = .00000104$, which indicates that the value given on the computer output should be interpreted as p-value $< .0001$. Because the value is so small, we can reject the hypothesis that tree growth retardation is not associated with soil pH. ■

EXAMPLE 11.6

The following data show the mean ages of executives of 15 firms in the food industry and the previous year's percentage increases in earnings per share of the firms. Use the Minitab output shown to test the hypothesis that executive age has no predictive value for change in earnings. Should a one-sided or two-sided alternative be used?

Mean age	x:	38.2	40.0	42.5	43.4	44.6	44.9	45.0	45.4
Change, earnings per share	y:	8.9	13.0	4.7	−2.4	12.5	18.4	6.6	13.5
	x:	46.0	47.3	47.3	48.0	49.1	50.5	51.6	
	y:	8.5	15.3	18.9	6.0	10.4	15.9	17.1	

```
Regression Analysis: ChgEarn versus MeanAge

Analysis of Variance

Source          DF   Adj SS  Adj MS  F-Value  P-Value
Regression       1   71.055  71.055     2.24    0.158
Error           13  412.602  31.739
Total           14  483.657

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
5.63371  14.69%      8.13%       0.00%

Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  -17.0     18.9    -0.90    0.384
MeanAge   0.617    0.413     1.50    0.158  1.00

Regression Equation:  ChgEarn = -17.0 + 0.617 MeanAge
```

Solution In the model $y = \beta_0 + \beta_1 x + \varepsilon$, the null hypothesis is H_0: $\beta_1 = 0$. The myth in American business is that younger managers tend to be more aggressive and harder driving, but it is also possible that the greater experience of the older executives leads to better decisions. Therefore, there is a good reason to choose a two-sided research hypothesis, H_a: $\beta_1 \neq 0$. The t statistic is shown in the output column marked T, reasonably enough. It shows $t = 1.50$, with a (two-sided) p-value of .158. There is not enough evidence to conclude that there is any relation between age and change in earnings.

In passing, note that the interpretation of $\hat{\beta}_0$ is rather interesting in this example; it would be the predicted change in earnings of a firm with the mean age of its managers equal to 0. Obviously, predictions should not be made at such small values for mean age. ■

It is also possible to calculate a confidence interval for the true slope. This is an excellent way to communicate the likely degree of inaccuracy in the estimate of that slope. The confidence interval once again is simply the estimate plus or minus a t table value times the standard error.

Confidence Interval for Slope β_1

$$\left(\hat{\beta}_1 - t_{\alpha/2}s_\varepsilon\sqrt{\frac{1}{S_{xx}}},\ \hat{\beta}_1 + t_{\alpha/2}s_\varepsilon\sqrt{\frac{1}{S_{xx}}}\right)$$

The required degrees of freedom for the table value $t_{\alpha/2}$ are $n - 2$, the error df.

EXAMPLE 11.7

Compute a 95% confidence interval for the slope β_1 using the output from Example 11.4.

Solution In the output, $\hat{\beta}_1 = -7.86$, and the estimated standard error of $\hat{\beta}_1$ is shown in the column labeled SE Coef as 1.09. Because n is 20, there are $20 - 2 = 18$ df for error. The required table value for $\alpha/2 = .05/2 = .025$ is 2.101. The corresponding confidence interval for the true value of β_1 is then

$$-7.86 \pm 2.101(1.09) \text{ or } -10.15 \text{ to } -5.57$$

The predicted decrease in growth retardation for a one-unit increase in soil pH ranges from -10.15 to -5.57. The large width of this interval is mainly due to the small sample size. ■

There is an alternative test, an F test, for the null hypothesis of no predictive value. It was designed to test the null hypothesis that *all* predictors have no value in predicting y. This test gives the same result as a two-sided t test of H_0: $\beta_1 = 0$ in simple linear regression; to say that all predictors have no value is to say that the (only) slope is 0. The F test is summarized next.

F Test for H_0: $\beta_1 = 0$

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

T.S.: $$F = \frac{\text{SS(Regression)}/1}{\text{SS(Error)}/(n-2)} = \frac{\text{MS(Regression)}}{\text{MS(Error)}}$$

R.R.: With $\text{df}_1 = 1$ and $\text{df}_2 = n - 2$, reject H_0 if $F > F_\alpha$.
Check assumptions and draw conclusions.

SS(Regression) is the sum of squared deviations of predicted y-values from the y mean. SS(Regression) $= \sum(\hat{y}_i - \bar{y}_i)^2$. SS(Error) is the sum of squared deviations of actual y-values from predicted y-values. SS(Error) $= \sum(\hat{y}_i - \bar{y}_i)^2$.

Virtually all computer packages calculate this F statistic. In Example 11.3, the output shows $F = 54.03$ with a p-value given by 0.000 (in fact, using R, p-value $= 1 -$ **pf**$(54.03, 1, 8) = .00008$). Again, the hypothesis of no predictive value can be rejected. It is always true for simple linear regression problems that $F = t^2$; in the example, $54.03 = (7.35)^2$, to within round-off error. The F and two-sided t tests are equivalent in simple linear regression; they serve different purposes in multiple regression.

EXAMPLE 11.8

For the output of Example 11.4, use the F-test for testing H_0: $\beta_1 = 0$. Show that $t^2 = F$ for this data set.

Solution The F statistic is shown in the output as 52.01, with a p-value of .000 (indicating the actual p-value is something less than .0005). Using a computer program, the actual p-value is .00000104. Note that the t statistic is -7.21 and $t^2 = (-7.21)^2 = 51.984$, which equals the F value, to within round-off error. ■

A confidence interval for β_0 can be computed using the estimated standard error of $\hat{\beta}_0$ as

$$\hat{\sigma}_{\hat{\beta}_0} = s_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

Confidence Interval for Intercept β_0

$$\hat{\beta}_0 \pm t_{\alpha/2}\, s_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

The required degrees of freedom for the table value of $t_{\alpha/2}$ are $n - 2$, the error df.

In practice, this parameter is of less interest than the slope. In particular, there is often no reason to hypothesize that the true intercept is zero (or any other particular value). Computer packages almost always test the null hypothesis of zero slope, but some don't bother with a test on the intercept term.

11.4 Predicting New *y*-Values Using Regression

In all the regression analyses we have done so far, we have been summarizing and making inferences about relations in data that have already been observed. Thus, we have been predicting the past. One of the most important uses of regression is trying to forecast the future. In the road-resurfacing example, the county highway director wants to predict the cost of a new contract that is up for bids. In a regression relating the change in systolic blood pressure for a specified dose of a drug, the doctor will want to predict the change in systolic blood pressure for a dose level not used in the study. In this section, we discuss how to make such regression predictions and how to determine prediction intervals that will convey our uncertainty in these predictions.

There are two possible interpretations of a y prediction based on a given x. Suppose that the highway director substitutes $x = 6$ miles in the regression equation $\hat{y} = 2.0 + 3.0x$ and gets $\hat{y} = 20$. This can be interpreted as either

> "The average cost $E(y)$ of *all* resurfacing contracts for 6 miles of road will be \$20,000."

or

> "The cost y of *this specific* resurfacing contract for 6 miles of road will be \$20,000."

The best-guess prediction in either case is 20, but the plus or minus factor differs. It is easier to estimate an average value $E(y)$ than predict an individual y-value, so the plus or minus factor should be less for estimating an average. We discuss the plus or minus range for estimating an average first, with the understanding that this is an intermediate step toward solving the specific-value problem.

In the mean-value estimating problem, suppose that the value of x is known. Because the previous values of x have been designated $x_1, \ldots, x_n$, call the new value x_{n+1}. Then $\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$ is used to predict $E(y_{n+1})$. Because $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, $\hat{y}_{n+1}$ is an unbiased predictor of $E(y_{n+1})$. The standard error of the estimated value can be shown to be

$$\sigma_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

Here S_{xx} is the sum of squared deviations of the original n values of x_i; it can be calculated from most computer outputs as

$$\left(\frac{s_\varepsilon}{\text{standard error } (\hat{\beta}_1)} \right)^2$$

Again, t tables with $n - 2$ df (the error df) must be used. The usual approach to forming a confidence interval—namely, estimate plus or minus t (standard error)—yields a confidence interval for $E(y_{n+1})$. Some of the better statistical computer packages will calculate this confidence interval if a new x-value is specified without specifying a corresponding y.

Confidence Interval for $E(Y_{n+1})$

$$\left(\hat{y}_{n+1} - t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}},\ \hat{y}_{n+1} + t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}} \right)$$

The degrees of freedom for the tabled t distribution are $n - 2$.

For the tree growth retardation study in Example 11.4, the computer output displayed here shows the estimated value of the average growth retardation, $E(y_{n+1})$, to be 16.0385 when the soil pH is $x = 4.0$. The corresponding 95% confidence interval on $E(y_{n+1})$ is 14.76 to 17.32.

```
Prediction for GrowthRet

Regression Equation:   GrowthRet = 47.48 - 7.86 SoilPh

Variable  Setting

SoilPh          4

    Fit    SE Fit         95% CI              95% PI
16.0385  0.609181   (14.7586, 17.3183)  (10.1791, 21.8979
```

The plus or minus term in the confidence interval for $E(y_{n+1})$ depends on the sample size n and the standard deviation around the regression line, as one

might expect. It also depends on the squared distance of x_{n+1} from $\bar{x}$ (the mean of the previous x_i values) relative to S_{xx}. As x_{n+1} gets farther from $\bar{x}$, the term

$$\frac{(x_{n+1} - \bar{x})^2}{S_{xx}}$$

gets larger. When x_{n+1} is far away from the other x-values, so that this term is large, the prediction is a considerable extrapolation from the data. Small errors in estimating the regression line are magnified by the extrapolation. The term $(x_{n+1} - \bar{x})^2/S_{xx}$ could be called an **extrapolation penalty** because it increases with the degree of extrapolation.

extrapolation penalty

Extrapolation—predicting the results at independent variable values far from the data—is often tempting and always dangerous. Using it requires an assumption that the relation will continue to be linear far beyond the data. By definition, you have no data to check this assumption. For example, a firm might find a negative correlation between the number of employees (ranging between 1,200 and 1,400) in a quarter and the profitability in that quarter: The fewer the employees, the greater the profit. It would be spectacularly risky to conclude from this fact that cutting the number of employees to 600 would vastly improve profitability. (Do you suppose we could have a negative number of employees?) Sooner or later, the declining number of employees must adversely affect the business, so that profitability turns downward. The extrapolation penalty term actually understates the risk of extrapolation. It is based on the assumption of a linear relation, and that assumption gets very shaky for large extrapolations.

The confidence and prediction intervals also depend heavily on the assumption of constant variance. In some regression situations, the variability around the line increases as the predicted value increases, violating this assumption. In such a case, the confidence and prediction intervals will be too wide where there is relatively little variability and too narrow where there is relatively large variability. A scatterplot that shows a "fan" shape indicates nonconstant variance. In such a case, the confidence and prediction intervals are not very accurate.

EXAMPLE 11.9

For the data of Example 11.4, and the following Minitab output from that data, obtain a 95% confidence interval for $E(y_{n+1})$ based on an assumed value for x_{n+1} of 6.5. Compare the width of the interval to one based on an assumed value for x_{n+1} of 4.0.

```
Prediction for GrowthRet

Regression Equation:   GrowthRet = 47.48 - 7.86 SoilPh

Variable  Setting

SoilPh          4

    Fit    SE Fit        95% CI              95% PI
16.0385  0.609181  (14.7586, 17.3183)  (10.1791, 21.8979)

Variable  Setting

SoilPh        6.5

     Fit   SE Fit         95% CI               95% PI
-3.60962  2.76491  (-9.41847, 2.19924)  (-11.7605, 4.54128)  XX

XX denotes an extremely unusual point relative to
predictor levels used to fit the model.
```

Solution For $x_{n+1} = 4.0$, the estimated value is equal to 16.0385. The confidence interval is shown as 14.7586 to 17.3183. For $x_{n+1} = 6.5$, the estimated value is -3.60962 with a confidence interval of -9.41847 to 2.19924. The second interval has a width 11.62, much larger than the first interval's width of 2.56. The value of $x_{n+1} = 6.5$ is far outside the range of x data; the extrapolation penalty makes the interval very wide compared to the width of intervals for values of x_{n+1} within the range of the observed x data. ■

prediction interval

Usually, the more relevant forecasting problem is that of predicting an individual y_{n+1} value rather than $E(y_{n+1})$. In most computer packages, the interval for predicting an individual value is called a **prediction interval.** The same best-guess $\hat{y}_{n+1}$ is used, but the forecasting plus or minus term is larger when predicting y_{n+1} than estimating $E(y_{n+1})$. In fact, it can be shown that the plus or minus forecasting error using $\hat{y}_{n+1}$ to predict y_{n+1} is as follows.

Prediction Interval for y_{n+1}

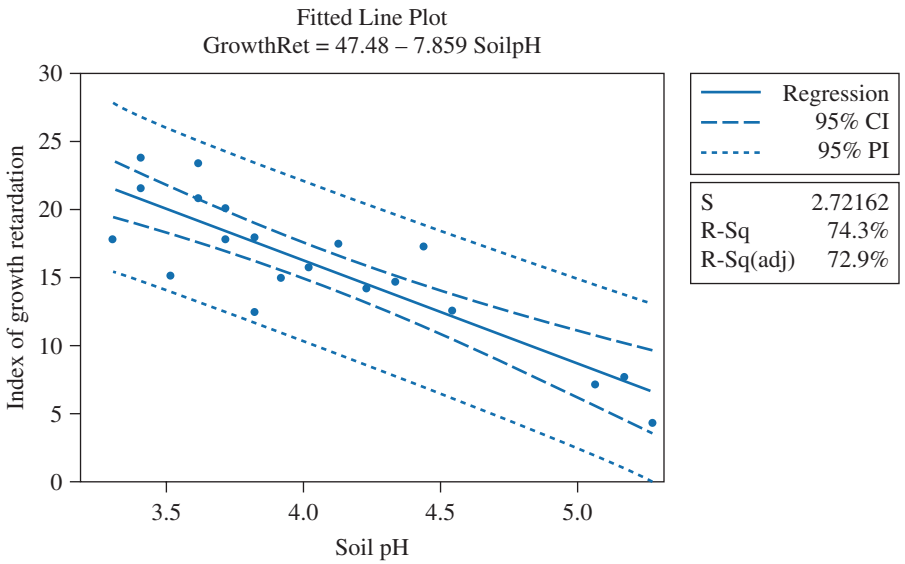
$$\left(\hat{y}_{n+1} - t_{\alpha/2}s_\varepsilon\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}},\ \hat{y}_{n+1} + t_{\alpha/2}s_\varepsilon\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}\right)$$

The degrees of freedom for the tabled t distribution are $n - 2$.

In the growth retardation example, the corresponding prediction limits for y_{n+1} when the soil pH $x = 4$ are 10.1791 to 21.8979 (see the output in Example 11.9). The 95% confidence intervals for $E(y_{n+1})$ and the 95% prediction intervals for y_{n+1} are plotted in Figure 11.14; the inner curves are for $E(y_{n+1})$, and the outer curves are for y_{n+1}.

The only difference between estimation of a mean $E(y_{n+1})$ and prediction of an individual y_{n+1} is the term $+1$ in the standard error formula. The presence of

FIGURE 11.14
Predicted values versus observed values with 95% prediction and confidence limits



this extra term indicates that predictions of individual values are less accurate than estimates of means. The extrapolation penalty term still applies, as does the warning that it understates the risk of extrapolation.

11.5 Examining Lack of Fit in Linear Regression

In our study of linear regression, we have been concerned with how well a linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ fits—but only from an intuitive standpoint. We could examine a scatterplot of the data to see whether it looked linear, and we could test whether the slope differed from 0; however, we had no way of testing to see whether a model containing terms such as $\beta_2 x^2, \beta_3 x^3$, and so on. would be a more appropriate model for the relationship between y and x. This section will outline situations in which we can test whether $y = \beta_0 + \beta_1 x + \varepsilon$ is an appropriate model.

Pictures (or graphs) are always a good starting point for examining lack of fit. First, use a scatterplot of y versus x. Second, a plot of residuals $y_i - \hat{y}_i$ versus predicted values $\hat{y}_i$ may give an indication of the following problems:

1. Outliers or erroneous observations. In examining the residual plot, your eye will naturally be drawn to data points with unusually high (in absolute value) residuals.
2. Violation of the assumptions. For the model $y = \beta_0 + \beta_1 x + \varepsilon$, we have assumed a linear relation between y and the dependent variable x, as well as independent, normally distributed errors with a constant variance.

The residual plot for a model and data set that has none of these apparent problems would look much like the plot in Figure 11.15. Note from this plot that there are no extremely large residuals (and hence no apparent outliers) and there is no trend in the residuals to indicate that the linear model is inappropriate. When a model containing terms such as $\beta_2 x^2, \beta_3 x^3$, and so on. is more appropriate, a residual plot more like that shown in Figure 11.16 would be observed.

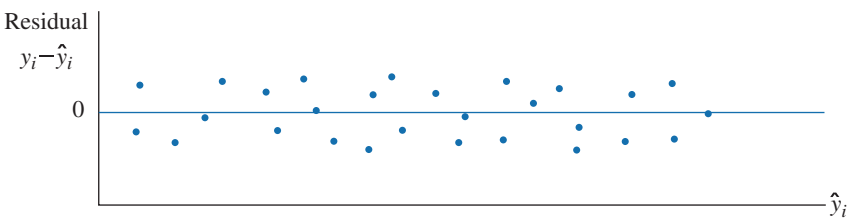


FIGURE 11.15
Residual plot with no apparent pattern

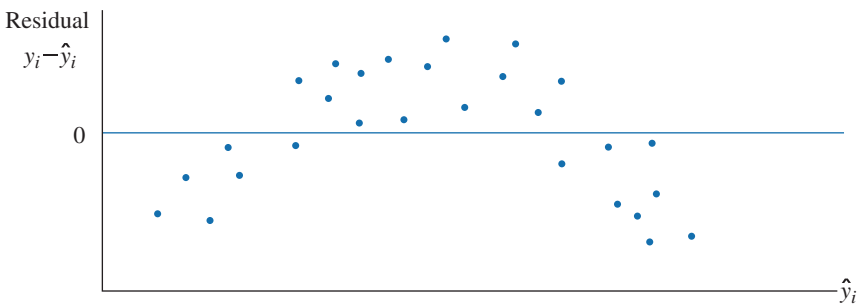


FIGURE 11.16
Residual plot showing the need for a higher-order model

FIGURE 11.17
Residual plot showing homogeneous error variances

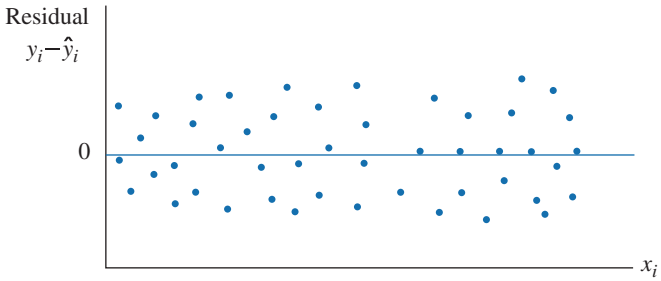
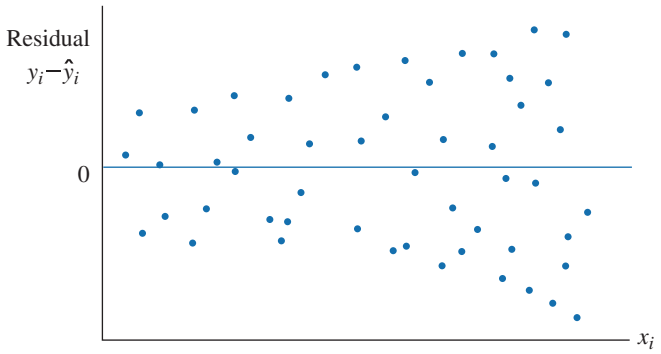


FIGURE 11.18
Residual plot showing error variances increasing with x



A check of the constant variance assumption can be addressed in the y versus x scatterplot or in a plot of the residuals $(y_i - \hat{y}_i)$ versus x_i. For example, a pattern of residuals as shown in Figure 11.17 indicates homogeneous error variances across values of x; Figure 11.18 indicates that the error variances increase with increasing values of x.

The question of independence of the errors and normality of the errors is addressed later in Chapter 13. We illustrate some of the points we have learned so far about residuals by way of an example.

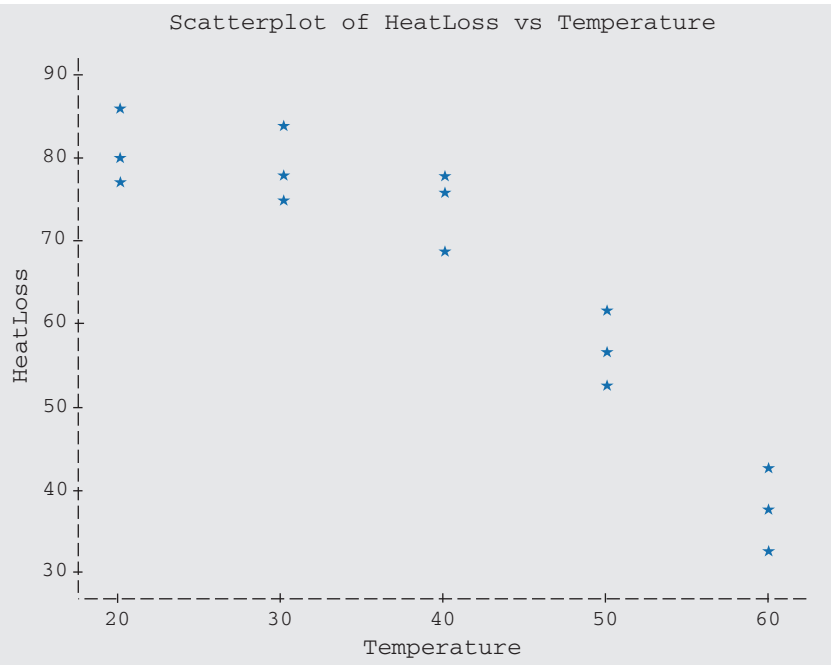
EXAMPLE 11.10

The manufacturer of a new brand of thermal panes examined the amount of heat loss by random assignment of three different panes to each of the three outdoor temperature settings being considered. For each trial, the window temperature was controlled at 68°F and 50% relative humidity.

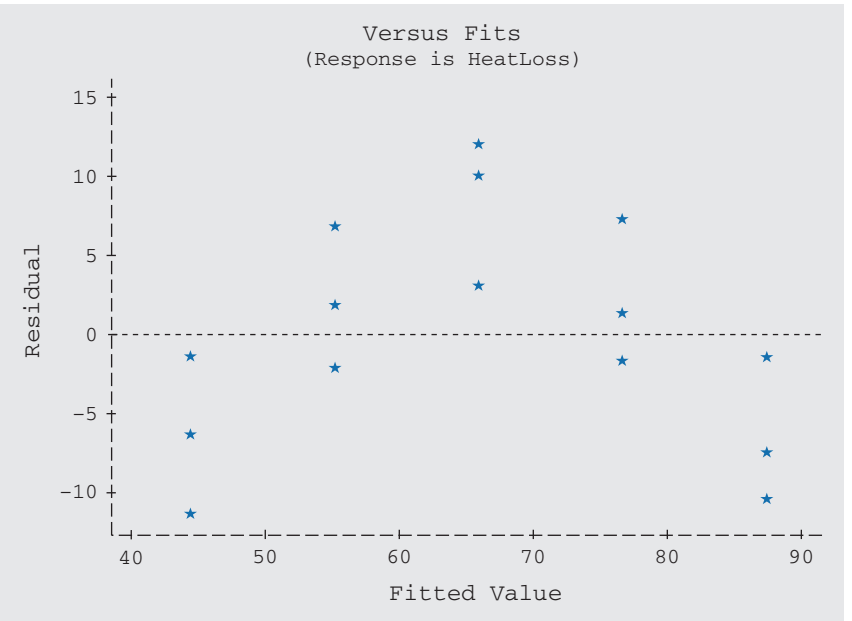
TABLE 11.4
Heat loss data

Outdoor Temperature (°F)	Heat Loss
20	86, 80, 77
30	78, 84, 75
40	78, 69, 76
50	62, 53, 57
60	33, 38, 43

a. Plot the data.
b. Fit the linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, and test $H_0: \beta_1 = 0$ (give the p-value for your test).




```
Regression Analysis: HeatLoss versus Temperature

Analysis of Variance

Source          DF  Adj SS   Adj MS  F-Value  P-Value
Regression       1  3477.6  3477.63    63.74    0.000
Error           13   709.3    54.56
  Lack-of-Fit    3   490.0   163.32     7.45    0.007
  Pure Error    10   219.3    21.93
Total           14  4186.9

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
7.38658  83.06%     81.76%      76.95%

Coefficients

Term           Coef  SE Coef  T-Value  P-Value   VIF
Constant     109.00     5.72    19.05    0.000
Temperature  -1.077    0.135    -7.98    0.000  1.00

Regression Equation:   HeatLoss = 109.00 - 1.077 Temperature

Fits and Diagnostics for All Observations

Obs  HeatLoss    Fit   Resid  Std Resid
  1     86.00  87.47   -1.47      -0.22
  2     80.00  87.47   -7.47      -1.13
  3     77.00  87.47  -10.47      -1.58
  4     78.00  76.70    1.30       0.19
  5     84.00  76.70    7.30       1.04
  6     75.00  76.70   -1.70      -0.24
  7     33.00  44.40  -11.40      -1.73
  8     38.00  44.40   -6.40      -0.97
  9     43.00  44.40   -1.40      -0.21
 10     78.00  65.93   12.07       1.69
 11     69.00  65.93    3.07       0.43
 12     76.00  65.93   10.07       1.41
 13     62.00  55.17    6.83       0.98
 14     53.00  55.17   -2.17      -0.31
 15     57.00  55.17    1.83       0.26
```



c. Compute $\hat{y}_i$ and $y_i - \hat{y}_i$ for the 15 observations. Plot $y_i - \hat{y}_i$ versus $\hat{y}_i$.
d. Does the constant variance assumption seem reasonable?

Solution The computer output shown here can be used to address the four parts of this example.

a. The scatterplot of heat loss versus temperature certainly shows a downward trend, and there is evidence of curvature as well.
b. The linear regression model seems to fit the data well, and the test of H_0: $\beta_1 = 0$ is significant with a p-value $= 1 - \mathbf{pt}(19.05, 13) < .0001$. However, is this the best model for the data?
c. The plot of residuals $(y_i - \hat{y}_i)$ against the fitted values $\hat{y}_i$ is similar to Figure 11.16, suggesting that we may need additional terms in our model.
d. It is clear from the original scatterplot and the residual plot that the constant variance condition appears to be valid. ■

pure experimental error
lack of fit

How can we test for the apparent lack of fit of the linear regression model in Example 11.10? When there is more than one observation per level of the independent variable, we can conduct a test for lack of fit of the fitted model by partitioning SS(Error) into two parts, one **pure experimental error** and the other **lack of fit.** Let y_{ij} denote the response for the jth observation at the ith level of the independent variable. Then, if there are n_i observations at the ith level of the independent variable, the quantity

$$\sum_j (y_{ij} - \bar{y}_i)^2$$

provides a measure of what we will call pure experimental error. This sum of squares has $n_i - 1$ degrees of freedom.

Similarly, for each of the other levels of x, we can compute a sum of squares due to pure experimental error. The pooled sum of squares

$$\text{SSP}_{\text{exp}} = \sum_{ij}(y_{ij} - \bar{y}_i)^2$$

called the sum of squares for pure experimental error, has $\sum_i(n_i - 1)$ degrees of freedom. With SS_{Lack} representing the remaining portion of SS(Error), we have

$$\text{SS(Error)} = \begin{matrix}\text{SSP}_{\text{exp}} \\ \text{due to pure} \\ \text{experimental} \\ \text{error}\end{matrix} + \begin{matrix}\text{SS}_{\text{Lack}} \\ \text{due to lack} \\ \text{to fit}\end{matrix}$$

If SS(Error) is based on $n - 2$ degrees of freedom in the linear regression model, then SS_{Lack} will have df $= n - 2 - \sum_i(n_i - 1)$.

Under the null hypothesis that our model is correct, we can form independent estimates of σ_ε^2, the model error variance, by dividing SSP_{exp} and SS_{Lack} by their respective degrees of freedom; these estimates are called **mean squares** and are denoted by MSP_{exp} and MS_{Lack}, respectively.

mean squares

The test for lack of fit is summarized here.

A Test for Lack of Fit in Linear Regression

H_0: A linear regression model is appropriate.
H_a: A linear regression model is not appropriate.

T.S.: $$F = \frac{\text{MS}_{\text{Lack}}}{\text{MSP}_{\text{exp}}}$$

where

$$\text{MS}_{\text{exp}} = \frac{\text{SSP}_{\text{exp}}}{\sum_i(n_i - 1)} = \frac{\sum_{ij}(y_{ij} - \bar{y}_i)^2}{\sum_i(n_i - 1)}$$

and

$$\text{MS}_{\text{Lack}} = \frac{\text{SS(Error)} - \text{SSP}_{\text{exp}}}{n - 2 - \sum(n_1 - 1)}$$

R.R.: For a specified value of α, reject H_0 (the adequacy of the model) if the computed value of F exceeds the table value for $\text{df}_1 = n - 2 - \sum_i(n_i - 1)$ and $\text{df}_2 = \sum_i(n_1 - 1)$.

Conclusion: If the F test is significant, this indicates that the linear regression model is inadequate. A nonsignificant result indicates that there is insufficient evidence to suggest that the linear regression model is inappropriate.

EXAMPLE 11.11

Refer to the data of Example 11.10. Conduct a test for lack of fit of the linear regression model.

Solution The contributions to experimental error for the differential levels of x are given in Table 11.5.

TABLE 11.5
Pure experimental error calculation

		Contribution to Pure Experimental Error	
Level of Temp.	$\bar{y}_i$	$\sum_{ij}(y_{ij} - \bar{y}_i)$	$n_i - 1$
20	81.00	42.00	2
30	79.00	42.00	2
40	74.33	44.66	2
50	57.33	40.66	2
60	38.00	50.00	2
Total		219.32	10

Summarizing these results, we have

$$\text{SSP}_{\text{exp}} = \sum_{ij}(y_{ij} - \bar{y}_i)^2 = 219.32 \qquad \text{with} \qquad \text{df}_2 = 10$$

The calculation of SSP_{exp} can be obtained by using the one-way ANOVA command in a software package. Using the theory from Chapter 8, designate the levels of the independent variable x as the levels of a treatment. The sum of squares error from this output is the value of SSP_{exp}. This concept is illustrated using the output from Minitab given here.

```
One-way ANOVA: HeatLoss versus Temperature

Factor          Levels         Values
Temperature        5     20, 30, 40, 50, 60

Analysis of Variance

Source       DF  Adj SS  Adj MS  F-Value  P-Value
Temperature   4  3967.6  991.90    45.22    0.000
Error        10   219.3   21.93
```

Note that the value of sum of squares error from the ANOVA is exactly the value that was computed above. Also, the degrees of freedom are given as 10, the same as in our calculations.

The output shown for Example 11.10 gives SS(Error) = 709.3; hence, by subtraction,

$$\text{SS}_{\text{Lack}} = \text{SS}(\text{Error}) - \text{SSP}_{\text{exp}} = 709.3 - 219.32 = 489.98$$

The sum of squares due to pure experimental error has $\sum_i(n_i - 1) = 10$ degrees of freedom; it therefore follows that with $n = 15$, SS_{Lack} has $n - 2 - \sum_i(n_i - 1) = 3$ degree of freedom. We find that

$$\text{MSP}_{\text{exp}} = \frac{\text{SSP}_{\text{exp}}}{10} = \frac{219.32}{10} = 21.93$$

and

$$\text{MS}_{\text{Lack}} = \frac{\text{SS}_{\text{Lack}}}{3} = \frac{489.98}{3} = 163.33$$

The F statistic for the test of lack of fit is

$$F = \frac{\text{MS}_{\text{Lack}}}{\text{MSP}_{\text{exp}}} = \frac{163.33}{21.93} = 7.45$$

Using $\text{df}_1 = 3$, $\text{df}_2 = 10$, and $\alpha = .05$, we will reject H_0 if $F \geq F_{.05,\,3,\,10} = 3.71$.

Because the computed value of F exceeds 3.71, we reject H_0 and conclude that there is significant lack of fit for a linear regression model with p-value = $1 - \textit{pf}(7.45, 3, 10) = .0066$. The scatterplot shown in Example 11.10 confirms that the model should be nonlinear in x.

The computer output from Example 11.10 confirms our calculations. ■

To summarize: In situations for which there is more than one y-value at one or more levels of x, it is possible to conduct a formal test for lack of fit of the linear regression model. This test should precede any inferences made using the fitted linear regression line. If the test for lack of fit is significant, some higher-order polynomial in x may be more appropriate. A scatterplot of the data and a residual plot from the linear regression line should help in selecting the appropriate model. More information on the selection of an appropriate model will be discussed along with multiple regression in Chapters 12 and 13.

If the F test for lack of fit is not significant, proceed with inferences based on the fitted linear regression line.

11.6 Correlation

Once we have found the prediction line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, we need to measure how well it predicts actual values. One way to do so is to look at the size of the residual standard deviation in the context of the problem. About 95% of the prediction errors will be within $\pm 2s_\varepsilon$. For example, suppose we are trying to predict the yield of a chemical process, where yields range from .50 to .94. If a regression model had a residual standard deviation of .01, we could predict most yields within $\pm .02$—fairly accurate in context. However, if the residual standard deviation was .08, we could predict most yields within $\pm .16$, which is not very impressive given that the yield range is only $.94 - .50 = .44$. This approach, though, requires that we know the context of the study well; an alternative, more general approach is based on the idea of correlation.

Suppose that we compare the squared prediction errors for two prediction methods: one using the regression model and the other ignoring the model and always predicting the mean y-value. In the road-resurfacing example of Section 11.2, if we are given the mileage values x_i, we could use the prediction equation $\hat{y}_i = 2.0 + 3.0x_i$ to predict costs. The deviations of actual values from predicted values, the residuals, measure prediction errors. These errors are summarized by the sum of squared residuals, $\text{SS(Error)} = \sum(y_i - \hat{y}_i)^2$, which is 44 for these data. For comparison, if we were not given the x_i values, the best squared error predictor of y would be the mean value $\bar{y} = 14$, and the sum of squared prediction errors would, in this case, be $\sum_i (y_i - \bar{y})^2 = \text{SS(Total)} = 224$. The proportionate reduction in error would be

$$\frac{\text{SS(Total)} - \text{SS(Error)}}{\text{SS(Total)}} = \frac{224 - 44}{224} = .804$$

In words, use of the regression model reduces squared prediction error by 80.4%, which indicates a fairly strong relation between the mileage to be resurfaced and the cost of resurfacing.

correlation coefficient

This proportionate reduction in error is closely related to the **correlation coefficient** of x and y. *A correlation measures the strength of the linear relation between* x *and* y. The stronger the correlation is, the better x predicts y, using $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Given n pairs of observations (x_i, y_i), we compute the sample correlation r as

$$r_{yx} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{S_{xx}\, S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}$$

where S_{xy} and S_{xx} are defined as before and

$$S_{yy} = \sum_i (y_i - \bar{y})^2 = \text{SS(Total)}$$

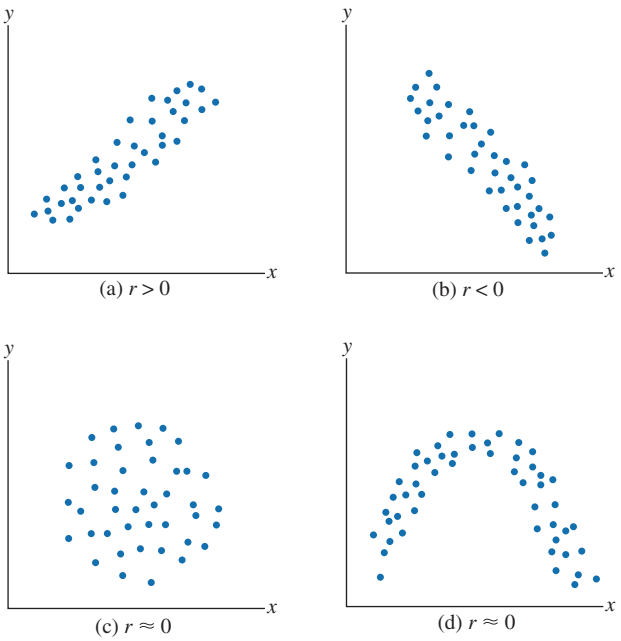
In the road-resurfacing example, $S_{xy} = 60$, $S_{xx} = 20$, and $S_{yy} = 224$, yielding

$$r_{yx} = \frac{60}{\sqrt{(20)(224)}} = .896$$

Generally, the correlation r_{yx} is a positive number if y tends to increase as x increases; r_{yx} is negative if y tends to decrease as x increases; and r_{yx} is zero if there is no relation between changes in x and changes in y or if there is a nonlinear relation such that patterns of increase and decrease in y (as x increases) cancel each other.

Figure 11.19 illustrates four possible situations for the values of r. In Figure 11.19(d), there is a strong relationship between y and x, but $r \approx 0$. This is a result of symmetric positive and negative nearly linear relationships canceling each

FIGURE 11.19
Interpretation of r



other. When $r = 0$, there is not a "linear" relationship between y and x. However, higher-order (nonlinear) relationships may exist. This situation illustrates the importance of plotting the data in a scatterplot. In Chapter 12, we will develop techniques for modeling nonlinear relationships between y and x.

EXAMPLE 11.12

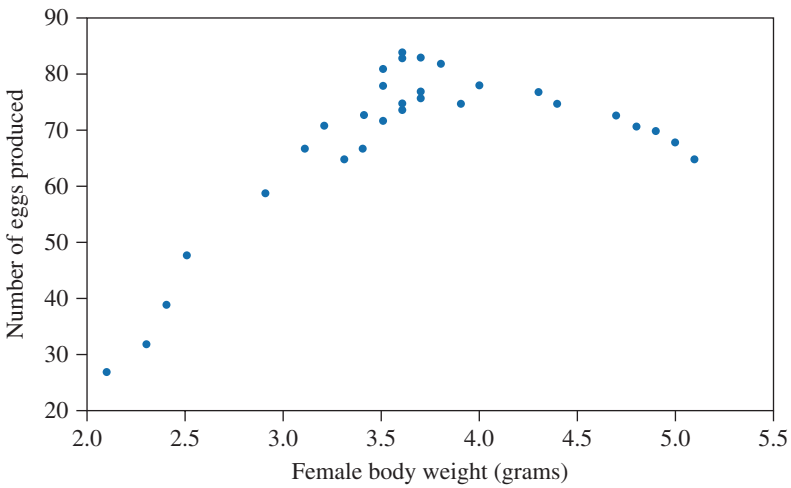
In a study of the reproductive success of grasshoppers, an entomologist collected a sample of 30 female grasshoppers. She recorded the number of mature eggs produced and the body weight of each of the females (in grams). The data are given here:

TABLE 11.6
Grasshopper egg data

Number of eggs, y	Weight of female, x (in grams)	Number of eggs, y	Weight of female, x (in grams)
27	2.1	75	3.6
32	2.3	84	3.6
39	2.4	77	3.7
48	2.5	83	3.7
59	2.9	76	3.7
67	3.1	82	3.8
71	3.2	75	3.9
65	3.3	78	4.0
73	3.4	77	4.3
67	3.4	75	4.4
78	3.5	73	4.7
72	3.5	71	4.8
81	3.5	70	4.9
74	3.6	68	5.0
83	3.6	65	5.1

A scatterplot of the data is displayed in Figure 11.20. Based on the scatterplot and an examination of the data, determine if the correlation should be positive or negative. Also, calculate the correlation between the number of eggs produced and the weight of the female.

FIGURE 11.20
Eggs produced versus female body weight



Solution Note that as the females' weight increases from 2.1 to 5.1, the number of eggs produced first increases and then for the last few females decreases. Therefore, the correlation is generally positive. Thus, we would expect the correlation coefficient to be a positive number.

The calculation of the correlation coefficient involves the same calculations needed to compute the least-squares estimates of the regression coefficients with one added sum of squares S_{xy}:

$$\sum_{i=1}^{30} x_i = 109.5 \Rightarrow \bar{x} = 3.65, \qquad \sum_{i=1}^{30} y_i = 2{,}605 \Rightarrow \bar{y} = 68.8333$$

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{30} (x_i - \bar{x})^2 \\ &= (2.1 - 3.65)^2 + (2.3 - 3.65)^2 + \cdots + (5.1 - 3.65)^2 = 17.615 \end{aligned}$$

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{30} (y_i - \bar{y})^2 \\ &= (27 - 68.8333)^2 + (32 - 68.8333)^2 + \cdots + (65 - 68.8333)^2 \\ &= 6{,}066.1667 \end{aligned}$$

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{30} (x_i - \bar{x})(y_i - \bar{y}) \\ &= (2.1 - 3.65)(27 - 68.8333) \\ &\quad + (2.3 - 3.65)(32 - 68.8333) + \cdots + (5.1 - 3.65)(65 - 68.8333) \\ &= 198.05 \end{aligned}$$

$$r_{xy} = \frac{198.05}{\sqrt{(17.615)(6{,}066.1667)}} = 0.606$$

The correlation is indeed a positive number. ■

coefficient of determination

Correlation and regression predictability are closely related. The proportionate reduction in error for regression we defined earlier is called the **coefficient of determination**. The coefficient of determination is simply the square of the correlation coefficient,

$$r_{yx}^2 = \frac{\text{SS(Total)} - \text{SS(Error)}}{\text{SS(Total)}}$$

which is the proportionate reduction in error. In the resurfacing example, $r_{yx} = .896$ and $r_{yx}^2 = .803$.

A correlation of zero indicates no predictive value in using the equation $y = \hat{\beta}_0 + \hat{\beta}_1 x$; that is, one can predict y as well without knowing x as one can knowing x. A correlation of 1 or -1 indicates perfect predictability—a 100% reduction in error attributable to knowledge of x. A correlation coefficient should routinely be interpreted in terms of its squared value, the coefficient of determination. Thus, a correlation of $-.3$, say, indicates only a 9% reduction in squared prediction error. Many books and most computer programs use the equation

$$\text{SS(Total)} = \text{SS(Error)} + \text{SS(Regression)}$$

where

$$\text{SS(Regression)} = \sum_i (\hat{y}_i - \bar{y})^2$$

Because the equation can be expressed as SS(Error) $= (1 - r_{yx}^2)$SS(Total), it follows that SS(Regression) $= r_{yx}^2$SS(Total), which again says that regression on x explains a proportion r_{yx}^2 of the total squared error of y.

EXAMPLE 11.13

For the grasshopper data in Example 11.12, compute SS(Total), SS(Regression), and SS(Error).

Solution SS(Total) $= S_{yy}$, which we computed to be 6,066.1667 in Example 11.13. We also found that $r_{yx} = 0.606$, so $r_{xy}^2 = (0.606)^2 = 0.367236$. Using the fact that SS(Regression) $= r_{yx}^2$ SS(Total), we have

$$\text{SS(Regression)} = (0.367236)\ (6{,}066.1667) = 2{,}227.7148.$$

From the equation SS(Error) = SS(Total) − SS(Regression), we obtain

$$\text{SS(Error)} = 6{,}066.1667 - 2{,}227.7148 = 3{,}838.45$$

Note that $r_{yx}^2 = (.606)^2 = 0.37$ indicates that a regression line predicting the number of eggs as a linear function of the weight of the female grasshopper would explain only about 37% of the variation in the number of eggs laid. This suggests that weight of the female is not a good predictor of the number of eggs. An examination of the scatterplot in Figure 11.20 shows a strong relationship between x and y, but the relationship is extremely nonlinear. A *linear* equation in x does not predict y very well, but a nonlinear equation would provide an excellent fit. ■

What values of r_{yx} indicate a "strong" relationship between y and x? Figure 11.21 displays 15 scatterplots obtained by randomly selecting 1,000 pairs (x_i, y_i) from 15 populations having bivariate normal distributions with correlations ranging from −0.99 to 0.99. We can observe that unless $|r_{yx}|$ is greater than 0.6 there is very little trend in the plot.

The sample correlation r_{yx} is the basis for estimation and significance testing of the population correlation ρ_{yx}. Statistical inferences are always based on assumptions. The assumptions of regression analysis—linear relation between x and y and constant variance around the regression line, in particular—are also assumed in **correlation inference**. In regression analysis, we regard the x-values as predetermined constants. In correlation analysis, we regard the x-values as randomly selected (and the regression inferences are conditional on the sampled x-values). If the xs are not drawn randomly, it is possible that the correlation estimates are biased. In some texts, the additional assumption is made that the x-values are drawn from a normal population. The inferences we make do not depend crucially on this normality assumption.

assumptions for correlation inference

The most basic inference problem is potential bias in the estimation of ρ_{yx}. A problem arises when the x-values are predetermined, as often happens in regression analysis. The choice of x-values can systematically increase or decrease the sample correlation. In general, a wide range of x-values tends to increase the magnitude of the correlation coefficient and a small range to decrease it. This effect is shown in Figure 11.22. If all the points in this scatterplot are included, there is an obvious,

FIGURE 11.21
Samples of size 1,000 from the bivariate normal distribution

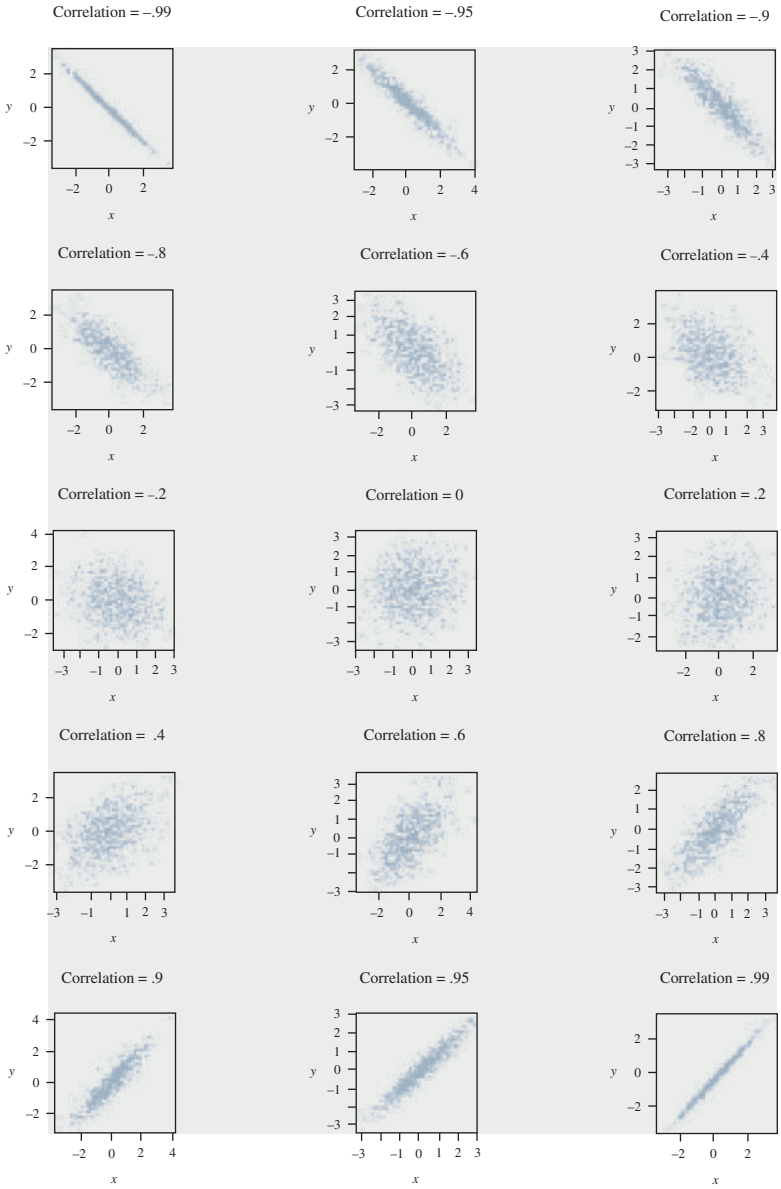
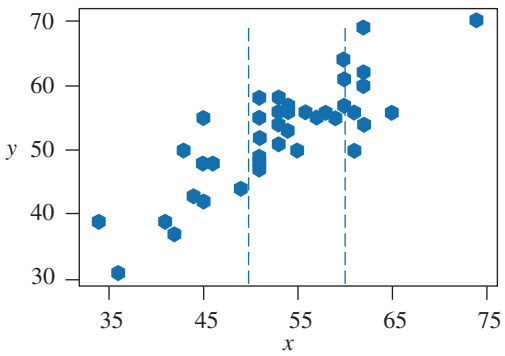


FIGURE 11.22
Effect of limited x range on sample correlation coefficient



strong correlation between x and y. Suppose, however, we consider only x-values in the range between the dashed vertical lines. By eliminating the outside parts of the scatter diagram, the sample correlation coefficient (and the coefficient of determination) are much smaller. Correlation coefficients can be affected by systematic choices of x-values; the residual standard deviation is *not* affected systematically, although it may change randomly if part of the x range changes. Thus, it is a good idea to consider the residual standard deviation s_ε and the magnitude of the slope when you decide how well a linear regression line predicts y.

EXAMPLE 11.14

The personnel director of a small company designs a study to evaluate the reliability of an aptitude test given to all newly hired employees. She randomly selects 12 employees that have been working for at least 1 year with the company and from their work records determines a productivity index (y) for each of the 12. The goal is to assess how strongly productivity correlates with the aptitude test (x).

y: 41 39 47 51 43 40 57 46 50 59 61 52
x: 24 30 33 35 36 36 37 37 38 40 43 49

Is the correlation larger or smaller if we consider only the six values with largest x-values?

```
                Simple Regression Analysis

Linear model: y = 20.5394 + 0.775176*x

                 Table of Estimates

                          Standard      t         P
             Estimate       Error     Value     Value
Intercept     20.5394     10.7251     1.92    0.0845
Slope        0.775176    0.289991     2.67    0.0234

R-squared = 41.68%
Correlation coeff. = 0.646
Standard error of estimation = 5.99236

File subset has been turned on, based on x>=37.

                Simple Regression Analysis

Linear model: y = 44.7439 + 0.231707*x

                 Table of Estimates

                          Standard      t         P
             Estimate       Error     Value     Value
Intercept     44.7439     24.8071     1.80    0.1456
Slope        0.231707    0.606677     0.38    0.7219

R-squared = 3.52%
Correlation coeff. = 0.188
Standard error of estimation = 6.34357
```

Solution For all 12 observations, the output shows a correlation coefficient of .646; the residual standard deviation is labeled as the standard error of estimation, 5.992. For the six highest x scores, shown as the subset having x greater than or equal to 37, the correlation is .188 and the residual standard deviation is 6.344. In

going from all 12 observations to the 6 observations with the highest x-values, the correlation has decreased drastically, but the residual standard deviation has hardly changed at all. ■

Just as we could run a statistical test for β_i, we can do it for ρ_{yx}.

Summary of a Statistical Test for ρ_{yx}

Hypotheses:

Case 1. H_0: $\rho_{yx} \leq 0$ versus H_a: $\rho_{yx} > 0$
Case 2. H_0: $\rho_{yx} \geq 0$ versus H_a: $\rho_{yx} < 0$
Case 3. H_0: $\rho_{yx} = 0$ versus H_a: $\rho_{yx} \neq 0$

T.S.: $t = r_{yx}\dfrac{\sqrt{n-2}}{\sqrt{1-r_{yx}^2}}$

R.R.: With $n - 2$ df and Type I error probability α,
1. $t > t_\alpha$.
2. $t < -t_\alpha$.
3. $|t| > t_{\alpha/2}$.

Check assumptions and draw conclusions.

We tested the hypothesis that the true slope is zero (in predicting tree growth retardation from soil pH) in Example 11.5; the resulting t statistic was -7.21. For those $n = 20$ stands, we can calculate r_{yx} as $-.862$ and r_{yx}^2 as .743. Hence, the correlation t statistic is

$$\frac{-.862\sqrt{18}}{\sqrt{1-.743}} = -7.21$$

An examination of the formulas for r and the slope $\hat{\beta}_1$ of the least-squares equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

yields the following relationship:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\sqrt{\frac{S_{yy}}{S_{xx}}} = r_{yx}\sqrt{\frac{S_{yy}}{S_{xx}}}$$

Thus, the t tests for a slope and for a correlation give identical results; it does not matter which form is used. It follows that the t test is valid for any choice of x-values. The bias we mentioned previously does not affect the sign of the correlation.

EXAMPLE 11.15

Perform t tests for the null hypothesis of zero correlation and zero slope for the data of Example 11.14 (all observations). Use an appropriate one-sided alternative.

Solution First, the appropriate H_a ought to be $\rho_{yx} > 0$ (and therefore $\beta_1 > 0$). It would be nice if an aptitude test had a positive correlation with the productivity score it was predicting! In Example 11.14, $n = 12$, $r_{yx} = .646$, and

$$t = \frac{.646\sqrt{12-2}}{\sqrt{1-(.646)^2}} = 2.68$$

Because this value falls between the tabled t-values for df = 10, $\alpha = .025$ (2.228) and for df = 10, $\alpha = .01$ (2.764), the p-value lies between .010 and .025. Hence, H_0 may be rejected. Using R, p-value = $2(1 - \mathbf{pt}(2.68, 10)) = .0231$.

The t statistic for testing the slope β_1 is shown in the output of Example 11.14 as 2.67, which equals (to within round-off error) the correlation t statistic, 2.68. The p-value = .0234. ■

The test for a correlation provides an interesting illustration of the difference between statistical significance and statistical importance. Suppose that a psychologist has devised a skills test for production-line workers and tests it on a huge sample of 40,000. If the sample correlation between test score and actual productivity is .02, then

$$t = \frac{.02\sqrt{39{,}998}}{\sqrt{1 - (.02)^2}} = 4.0$$

We would reject the null hypothesis at any reasonable α level, so the correlation is "statistically significant." However, the test accounts for only $(.02)^2 = .0004$ of the squared error in skill scores, so it is *almost* worthless as a predictor. Remember, the rejection of the null hypothesis in a statistical test is the conclusion that the sample results cannot plausibly have occurred by chance if the null hypothesis is true. The test itself does not address the practical significance of the result. Clearly, for a sample size of 40,000, even a trivial sample correlation like .02 is not likely to occur by mere luck of the draw. However, there is no practically meaningful relationship between these test scores and productivity scores in this example.

In most situations, it is also of interest to obtain confidence limits on ρ_{yx} to assess the uncertainty in its estimation when using the sample correlation coefficient, r_{yx}.

Confidence Interval for the Correlation Coefficient ρ_{yx}

A $100(1 - \alpha/2)$ confidence interval for ρ_{yx} is given by

$$\left(\frac{e^{2z_1} - 1}{e^{2z_1} + 1}, \frac{e^{2z_2} - 1}{e^{2z_2} + 1}\right)$$

where

$$z = \frac{1}{2}\ln\left(\frac{1 + r_{yx}}{1 - r_{yx}}\right)$$

$$z_1 = z - \frac{z_{\alpha/2}}{\sqrt{n - 3}}$$

$$z_2 = z + \frac{z_{\alpha/2}}{\sqrt{n - 3}}$$

and $z_{\alpha/2}$ is obtained from Table 1 in the Appendix.

The above confidence interval requires that the n pairs (x_i, y_i) have a bivariate normal distribution or that n is fairly large.

EXAMPLE 11.16

Use the data in Example 11.12 to place a 95% confidence interval on the correlation between the number of mature eggs and the weight of the female grasshopper.

Solution From the data in Example 11.12, $n = 30$, $r_{yx} = .606$, and the value of $z_{\alpha/2} = z_{.025} = 1.96$. Next, compute Fisher's transformation of r_{yx}:

$$z = \frac{1}{2}\ln\left(\frac{1 + r_{yx}}{1 - r_{yx}}\right) = \frac{1}{2}\ln\left(\frac{1 + .606}{1 - .606}\right) = .70258$$

$$z_1 = z - \frac{z_{\alpha/2}}{\sqrt{n - 3}} = .70258 - \frac{1.96}{\sqrt{30 - 3}} = .32538$$

$$z_2 = z + \frac{z_{\alpha/2}}{\sqrt{n - 3}} = .70258 + \frac{1.96}{\sqrt{30 - 3}} = 1.07978$$

The 95% confidence interval for ρ_{yx} is given by

$$\left(\frac{e^{2z_1} - 1}{e^{2z_1} + 1}, \frac{e^{2z_2} - 1}{e^{2z_2} + 1}\right) = \left(\frac{e^{2(.32538)} - 1}{e^{2(.32538)} + 1}, \frac{e^{2(1.07978)} - 1}{e^{2(1.07978)} + 1}\right) = (.314, .793)$$

With 95% confidence, we would estimate that the correlation coefficient is between .314 and .793, whereas the point estimator r_{yx} was given as .606. The width of the 95% confidence interval reflects the uncertainty in using r_{yx} as an estimator of the correlation coefficient when the sample size is small. ■

Spearman rank correlation coefficient

The correlation coefficient, r_{yx}, assesses the linear association between two variables x and y. In some circumstances, one or both of these variables will not be numerical but will be ordinal, in which case the value of r_{yx} cannot be computed. In other cases, the distribution of the x and y may be highly skewed—that is, very nonnormal in distribution. In both of these situations, the significance of the correlation cannot be assessed using r_{xy}. An approach for assessing *monotonic* association between two variables is to use the **Spearman rank correlation coefficient,** r_s. The rank correlation measures whether y increases (or decreases) with increases in x, even in those situations where the relation between y and x is not necessarily linear.

The Spearman rank correlation coefficient is computed by first ranking the values of x and the values of y and then computing the ordinary correlation coefficient for the data set consisting of the ranks.

EXAMPLE 11.17

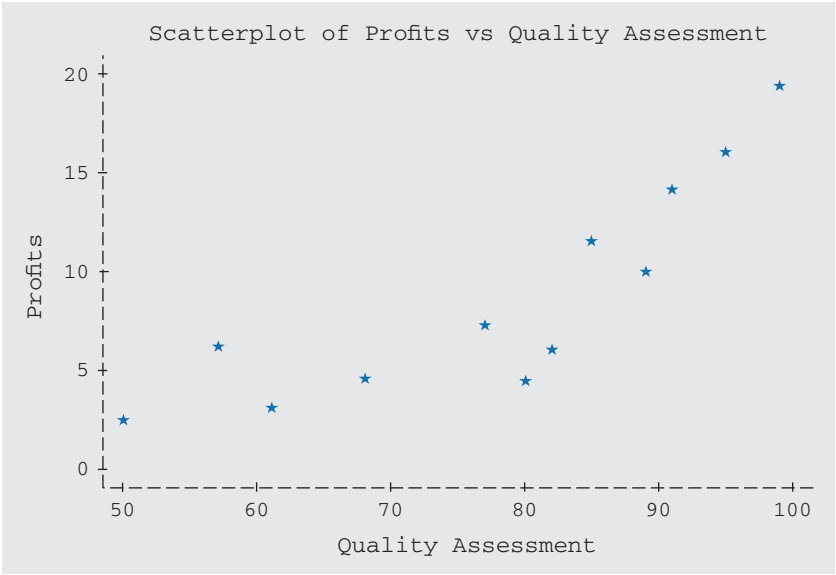
A corporation examined the relationship between the profit for its 12 product lines ($10,000) and the overall quality assessment of the 12 products (scale of 0 to 100). The data is given in the following table.

Profit	2.5	6.2	3.1	4.6	7.3	4.5	6.1	11.6	10.0	14.2	16.1	19.5
Quality assessment	50	57	61	68	77	80	82	85	89	91	95	99

a. Plot the data. Is profit linearly related to quality?
b. Compute the Spearman rank correlation coefficient.

Solution

a. A plot of the data reveals that as quality increases there is a general increase in profit, but it is not linear.



b. To compute Spearman's rank correlation coefficient, r_s, first we need to replace the data values with their ranks, determined separately for each of the variables.

Profit ranks	1	6	2	4	7	3	5	9	8	10	11	12
Quality assessment ranks	1	2	3	4	5	6	7	8	9	10	11	12

Let y denote the ranks on profits and x denote the ranks on quality assessment. The computation of r_s follows the same steps as we used in computing the ordinary correlation coefficient, r. Thus, we need to compute S_{xx}, S_{yy}, and S_{xx}, yielding the following values:

$$S_{xx} = \sum_{i=1}^{12}(x_i - \bar{x})^2 = \sum_{i=1}^{12}(x_i - 6.5)^2 = 143$$

$$S_{yy} = \sum_{i=1}^{12}(y_i - \bar{y})^2 = \sum_{i=1}^{12}(y_i - 6.5)^2 = 143$$

$$S_{xy} = \sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{12}(x_i - 6.5)(y_i - 6.5) = 125$$

Hence, the Spearman rank correlation coefficient is computed as follows.

$$r_s = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} = \frac{125}{\sqrt{(143)(143)}} = .874$$

The values of the Spearman rank correlation coefficient range from −1 to 1. The value $r_s = .874$ would indicate a strong relationship between profit and quality of the product. ■

11.7 RESEARCH STUDY: Two Methods for Detecting *E. coli*

The research study in Chapter 7 described a new microbial method for the detection of *E. coli*, the Petrifilm HEC test. The researchers wanted to evaluate the agreement of the results obtained using the HEC test with the results obtained from an elaborate laboratory-based procedure, hydrophobic grid membrane filtration (HGMF). The HEC test is easier to inoculate, more compact to incubate, and safer to handle than conventional procedures. However, prior to using the HEC procedure, it was necessary to compare the readings from the HEC test to the readings from the HGMF procedure obtained on the same meat sample. This would determine whether the two procedures were yielding essentially the same readings. If the readings differed but an equation could be obtained that could closely relate the HEC reading to the HGMF reading, then the researchers could calibrate the HEC readings to predict what readings would have been obtained using the HGMF test procedure. If the HEC test results were unrelated to the HGMF test procedure results, then the HEC test could not be used in the field in detecting *E. coli*.

Designing Data Collection

We described in Chapter 7 Phase One of the experiment. Phase Two of the study was to apply both procedures to artificially contaminated beef. Portions of beef trim were obtained from three Holstein cows that had tested negative for *E. coli*. Eighteen portions of beef trim were obtained from the cows and then contaminated with *E. coli.* The HEC and HGMF procedures were applied to a portion of each of the 18 samples. The two procedures yielded *E. coli* concentrations in transformed metric ($\log_{10}$ CFU/ml). The data consisted of 18 pairs of observations and are given in Table 11.7.

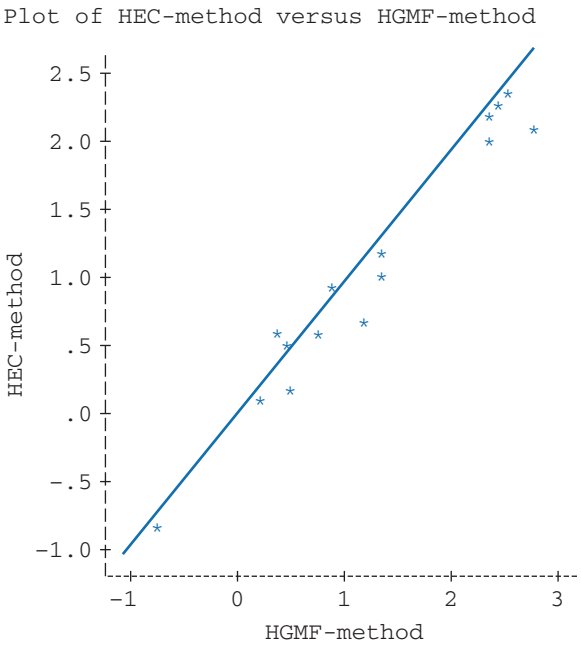
Managing the Data

The researchers would next prepare the data for a statistical analysis following the steps described in Section 2.5 of the textbook. They would need to carefully

TABLE 11.7
Data for research study

RUN	HEC	HGMF	RUN	HEC	HGMF
1	.50	.42	10	1.20	1.25
2	.06	.20	11	.93	.83
3	.20	.42	12	2.27	2.37
4	.61	.33	13	2.02	2.21
5	.20	.42	14	2.32	2.44
6	.56	.64	15	2.14	2.28
7	−.82	−.82	16	2.09	2.69
8	.67	1.06	17	2.30	2.43
9	1.02	1.21	18	−.10	1.07

FIGURE 11.23
Plot of HEC method versus HGMF method



NOTE: Two obs hidden.

review the experimental procedures to make sure that each pair of meat samples was nearly identical so as not to introduce any differences in the HEC and HGMF readings that were not part of the differences in the two procedures. During such a review, procedural problems during run 18 were discovered, and this pair of observations was excluded from the analysis.

Analyzing the Data

The researchers were interested in determining if the two procedures yielded measures of *E. coli* concentrations that were strongly related. The scatterplot of the experimental data is given in Figure 11.23.

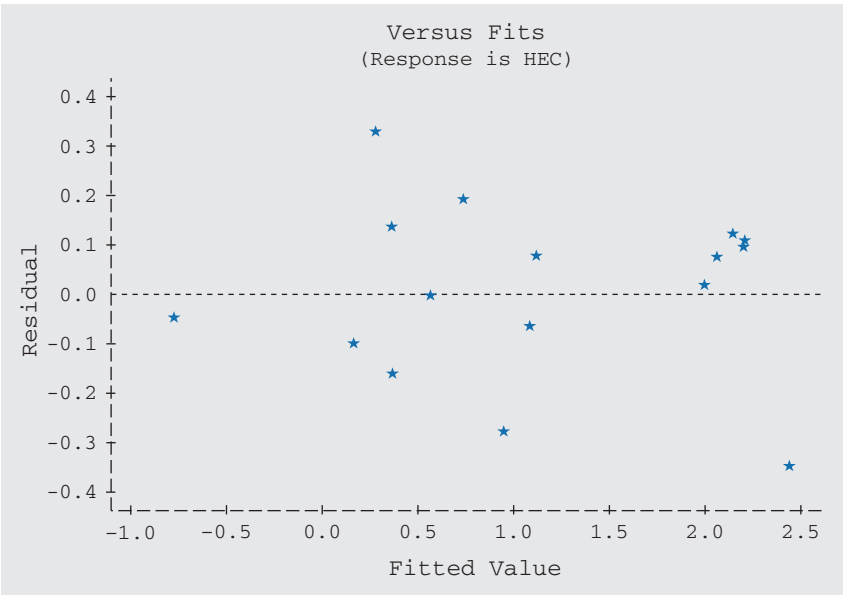
A 45° line was placed in the scatterplot to display the relative agreement between the readings from the two procedures. If the plotted points fell on this line, then the two procedures would be in complete agreement in their determination of *E. coli* concentrations. Although the 17 points are obviously highly correlated, they are not equally scattered about the 45° line; 14 of the points are below the line, with only three points above the line. Thus, the researchers would like to determine an equation that would relate the readings from the two procedures. If the two readings from the two procedures can be accurately related using a regression equation, the researchers would be able to predict the reading of the HGMF procedure given the HEC reading on a meat sample. This would enable them to compare *E. coli* concentrations obtained from meat samples in the field using the HEC procedure to the readings obtained in the laboratory using the HGMF procedure.

The researchers were interested in assessing the degree to which the HEC and HGMF procedures agreed in determining the level of *E. coli* concentrations in meat samples. We will first obtain the regression relationship, with HEC serving as the dependent variable and HGMF as the independent variable, since the HGMF procedure has a known reliability in determining *E. coli* concentrations.

The computer output for analyzing the 17 pairs of *E. coli* concentrations is given here along with a plot of the residuals.

```
Dependent Variable: HEC                          HEC-METHOD

Analysis of Variance

                          Sum of          Mean
Source          DF       Squares        Square      F Value        Prob>F

Model            1      14.22159      14.22159      441.816        0.0001
Error           15       0.48283       0.03219
C Total         16      14.70442


    Root MSE       0.17941     R-square       0.9672
    Dep Mean       1.07471     Adj R-sq       0.9650
    C.V.          16.69413
```

```
                    Parameter       Standard    T for H0:
 Variable  DF       Estimate          Error    Parameter=0    Prob > |T|

 INTERCEP   1      -0.023039     0.06797755         -0.339        0.7394
 HGMF       1       0.915685     0.04356377         21.019        0.0001
```



The R^2 value of .9672 indicates a strong linear relationship between HEC and HGMF concentrations. An examination of the residual plots does not indicate the necessity for higher-order terms in the model or for heterogeneity in the variances. The least-squares equation relating HEC to HGMF concentrations is given here.

$$\widehat{\text{HEC}} = -.023 + .9157 * \text{HGMF}$$

Thus, we can assess whether there is an exact relationship between the two methods of determining *E. coli* concentrations by testing the hypotheses

$$H_0: \beta_0 = 0 \text{ and } \beta_1 = 1 \text{ versus } H_a: \beta_0 \neq 0 \text{ and/or } \beta_1 \neq 1$$

If H_0 were accepted, then we would have a strong indication that the relationship $\widehat{\text{HEC}} = 0 + 1 * \text{HGMF}$ was valid. That is, HEC and HGMF were yielding essentially the same values for *E. coli* concentrations. From the output, we have a p-value = .7394 for testing H_0: $\beta_0 = 0$, and we can test H_0: $\beta_1 = 1$ using the test statistic

$$t = \frac{\hat{\beta}_1 - 1}{\widehat{SE}(\hat{\beta}_1)} = \frac{.915685 - 1}{.04356377} = -1.935$$

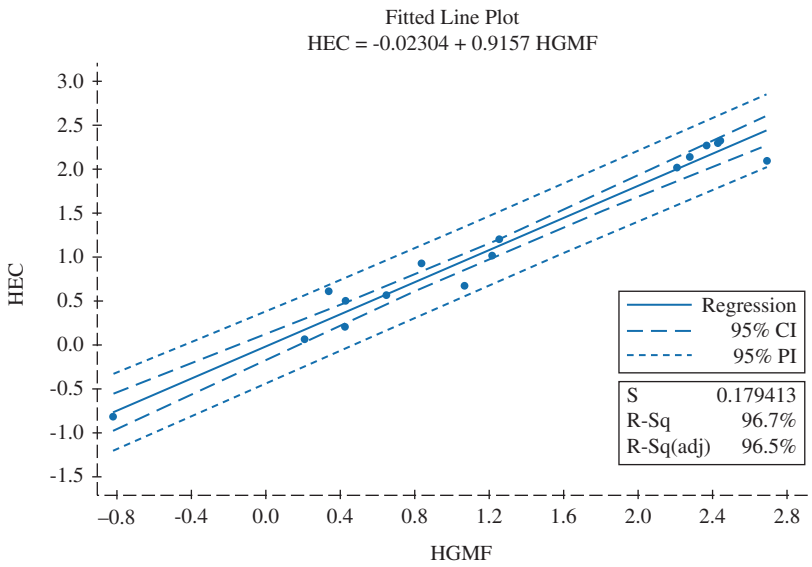
The p-value of the test statistic is p-value = $Pr(|t_{15}| \geq 1.935) = .0721$. In order to obtain an overall α value of .05, we evaluate the hypotheses of H_0: $\beta_0 = 0$ and H_0: $\beta_1 = 1$ individually using $\alpha = .025$; that is, we reject an individual hypothesis if its p-value is less than .025. Because the p-values are .7394 and .0721, we fail to reject either null hypothesis and conclude that the data do not support the hypothesis that HEC and HGMF are yielding significantly different *E. coli* concentrations.

Even though HEC and HGMF are not yielding exactly the same determinations, by solving the regression equation for HGMF in terms of HEC, the value of HGMF could be predicted from the value of HEC:

$$\widehat{HGMF} = (HEC + .023)/.9157 = .025 + 1.092HEC$$

Figure 11.24 contains the regression equation relating HEC to HGMF along with 95% confidence and prediction lines. Using the prediction lines, a 95% prediction interval can be determined for the predicted value of HGMF for a given value of HEC. The procedure involves drawing a horizontal line at the level of the specified value of HEC. Next, the intersections of the the horizontal line with the 95% prediction lines are projected to the HGMF axis. The two points on the HGMF axis would be the 95% prediction interval for HGMF for the given value of HEC. For example, if HEC = .5, then the corresponding values on the HGMF axis are .16 and 1.04. We can then conclude that when HEC = .5, a 95% prediction interval for the values of HGMF would be (.16, 1.04).

FIGURE 11.24
Plot of regression of HEC on HGMF



11.8 Summary and Key Formulas

This chapter introduces regression analysis and is devoted to simple regression, using only one independent variable to predict a dependent variable. The basic questions involve the nature of the relation (linear or curved), the amount of variability around the predicted value, whether that variability is constant over the range of prediction, how useful the independent variable is in predicting the dependent variable, and how much to allow for sampling error. The key concepts of the chapter include the following:

1. The data should be plotted in a scatterplot. A smoother such as LOWESS or a spline curve is useful in deciding whether a relation is nearly linear or is clearly curved. Curved relations can often be made nearly linear by transforming either the independent variable or the dependent variable or both.
2. The coefficients of a linear regression are estimated by least squares, which minimizes the sum of squared residuals (actual values minus predicted values). Because squared error is involved, this method is sensitive to outliers.
3. Observations that are extreme in the x (independent variable) direction have high leverage in fitting the line. If a high leverage point also falls well off the line, it has high influence, in that removing the observation substantially changes the fitted line. A high influence point should be omitted if it comes from a different population than the remainder. If it must be kept in the data, a method other than least squares should be considered.
4. Variability around the line is measured by the standard deviation of the residuals. This residual standard deviation may be interpreted using the Empirical Rule. The residual standard deviation sometimes increases as the predicted value increases. In such a case, try transforming the dependent variable.
5. Hypothesis tests and confidence intervals for the slope of the line (and, less interestingly, the intercept) are based on the t distribution. If there is no relation, the slope is 0. The line is estimated most accurately if there is a wide range of variation in the x-variable.
6. The fitted line may be used to forecast at a new x-value, again using the t distribution. This forecasting is potentially inaccurate if the new x-value is extrapolated beyond the support of the observed data.
7. A standard method of measuring the strength of relation is the coefficient of determination, the square of the correlation. This measure is diminished by nonlinearity or by an artificially limited range of x variation.

One of the most important uses of statistics for managers is prediction. A manager may want to forecast the cost of a particular contracting job given the size of that job, to forecast the sales of a particular product given the current rate of growth of the gross national product, or to forecast the number of parts that will be produced given a certain size workforce. The statistical method most widely used in making predictions is *regression analysis.*

In the regression approach, past data on the relevant variables are used to develop and evaluate a prediction equation. The variable that is being predicted

by this equation is the dependent variable. A variable that is being used to make the prediction is an independent variable. In this chapter, we discuss regression methods involving a single independent variable. In Chapter 12, we extend these methods to multiple regression, the case of several independent variables.

A number of tasks can be accomplished in a regression study:

1. The data can be used to obtain a prediction equation.
2. The data can be used to estimate the amount of variability or uncertainty around the equation.
3. The data can be used to identify unusual points far from the predicted value, which may represent unusual problems or opportunities.
4. Because the data are only a sample, inferences can be made about the true (population) values for the regression quantities.
5. The prediction equation can be used to predict a reasonable range of values for future values of the dependent variable.
6. The data can be used to estimate the degree of correlation between de-pendent and independent variables, a measure that indicates how strong the relation is.

Key Formulas

1. Least-squares estimates of slope and intercept

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

and

$$S_{xx} = \sum_i (x_i - \bar{x})^2$$

2. Estimate of σ_ε^2

$$s_\varepsilon^2 = \frac{\Sigma_i (y_i - \hat{y}_i)^2}{n - 2} = \frac{\text{SS(Error)}}{n - 2}$$

3. Statistical test for β_1

H_0: $\beta_1 = 0$ (two-tailed)

$$\text{T.S.: } t = \frac{\hat{\beta}_1}{s_\varepsilon / \sqrt{S_{xx}}}$$

4. Confidence interval for β_1

$$\hat{\beta}_1 \pm t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{S_{xx}}}$$

5. F test for β_1

H_0: $\beta_1 = 0$ (two-tailed)

$$\text{T.S.: } F = \frac{\text{MS(Regression)}}{\text{MS(Error)}}$$

6. Confidence interval for $E(y_{n+1})$

$$\hat{y}_{n+1} \pm t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

7. Prediction interval for y_{n+1}

$$\hat{y}_{n+1} \pm t_{\alpha/2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

8. Test for lack of fit in linear regression

$$\text{T.S.: } F = \frac{\text{MS}_{\text{Lack}}}{\text{MSP}_{\text{exp}}}$$

where

$$\text{MSP}_{\text{exp}} = \frac{\text{SSP}_{\text{exp}}}{\Sigma_i (n_i - 1)} = \frac{\Sigma_{ij} (y_{ij} - \bar{y}_i)^2}{\Sigma_i (n_i - 1)}$$

and

$$MS_{Lack} = \frac{SS(Error) - SSP_{exp}}{(n-2) - \sum_i (n_i - 1)}$$

9. Correlation coefficient

$$r_{yx} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}}$$

10. Coefficient of determination

$$r_{yx}^2 = \frac{SS(Total) - SS(Error)}{SS(Total)}$$

11. Confidence interval for ρ_{yx}

$$\left(\frac{e^{2z_1} - 1}{e^{2z_1} + 1}, \frac{e^{2z_2} - 1}{e^{2z_2} + 1}\right)$$

12. Statistical test for ρ_{yx}

H_0: $\rho_{yx} = 0$ (two-tailed)

$$\text{T.S.: } t = r_{yx}\frac{\sqrt{n-2}}{\sqrt{1 - r_{yx}^2}}$$

13. Spearman rank correlation coefficient

$$r_s = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}}$$

where x and y are ranks

11.9 Exercises

11.2 Estimating Model Parameters

Basic **11.1** Plot the data shown here in a scatterplot, and sketch a line through the points.

x	4	9	14	19	24	29	34	39	43
y	4	18	23	22	37	38	47	50	64

Basic **11.2** Refer to Exercise 11.1.

a. Plot the equation $\hat{y} = .51 + 1.38x$ in the scatterplot produced in Exercise 11.1. Comment on how close this line is to the line you fitted through the points.

b. Use the equation $\hat{y} = .51 + 1.38x$ to predict y for $x = 20$.

Basic **11.3** Use the data given here to answer the following questions.

x	7	12	14	22	27	33	37	39	42	49	53	61
y	10.6	16.8	23.3	12.5	91.7	67.7	130.7	110.3	147.3	138.3	142.6	151.4

a. Plot the data values in a scatter diagram.

b. Sketch a straight line through the points.

c. Use your sketched line to predict the value of y when $x = 40$.

Basic **11.4** Refer to Exercise 11.3.

a. Determine the least-squares prediction equation.

b. Use the least-squares prediction equation to predict y when $x = 40$.

c. Compare your prediction from part (b) to your prediction from Exercise 11.3.

Basic **11.5** Refer to the Exercise 11.4.

a. Use the least-squares prediction equation to predict y when $x = 100$.

b. Comment on the validity of this prediction.

Basic **11.6** Use the output from Minitab for these data to answer the following questions.

x	20	36	50	80	95	121	85	63	98	108
y	32	75	87	152	195	274	184	123	136	203

a. Plot the data on a scatterplot.

b. Locate the least-squares prediction from the output given here, and draw the regression line in the scatterplot.

c. Does the predicted equation seem to represent the data adequately?
d. Predict y when $x = 77$.

```
Minitab Output:

Regression Analysis: y versus x

Analysis of Variance

Source      DF  Adj SS   Adj MS  F-Value  P-Value
Regression   1   40627  40626.9    64.56    0.000
  x          1   40627  40626.9    64.56    0.000
Error        8    5034    629.3
Total        9   45661

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
25.0850  88.98%     87.60%      82.69%

Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant   -9.7     20.9    -0.46    0.657
x         2.060    0.256     8.04    0.000  1.00


Regression Equation:  y = -9.7 +2.060x

Fits and Diagnostics for All Observations

Obs      y    Fit  Resid  Std Resid
  1   32.0   31.5    0.5       0.02
  2   75.0   64.5   10.5       0.49
  3   87.0   93.4   -6.4      -0.28
  4  152.0  155.2   -3.2      -0.13
  5  195.0  186.1    8.9       0.38
  6  274.0  239.6   34.4       1.66
  7  184.0  165.5   18.5       0.78
  8  123.0  120.1    2.9       0.12
  9  136.0  192.3  -56.3      -2.44  R
 10  203.0  212.9   -9.9      -0.44

R  Large residual
```

Ag. **11.7** A food processor was receiving complaints from its customers about the firmness of its canned sweet potatoes. The firm's research scientist decided to conduct an experiment to determine if adding pectin to the sweet potatoes might result in a product with a more desirable firmness. The experiment was designed using three concentrations of pectin (by weight)—1%, 2%, and 3%—and a control with 0%. The processor packed 12 cans with sweet potatoes with a 25% (by weight) sugar solution. Three cans were randomly assigned to each of the pectin concentrations with the appropriate percentage of pectin added to the sugar syrup. The cans were sealed and placed in a 25°C environment for 30 days. At the end of the storage time, the cans were opened, and a firmness determination was made for the contents of each can. These appear below:

Pectin concentration	0%	1%	2%	3%
Firmness reading	46.90, 50.20, 51.30	56.48, 59.34, 62.97	67.91, 70.78, 73.67	68.13, 70.85, 72.34

a. Let x denote the pectin concentration of the sweet potatoes in a can and y denote the firmness reading following the 30 days of storage at 25°C. Plot the sample data in a scatter diagram.

b. Obtain the least-squares estimates for the parameters, and plot the least-squares line on your scatter diagram.
c. Does firmness appear to be in a constant increasing relation with pectin concentration?
d. Predict the firmness of a can of sweet potatoes treated with a 1.5% concentration of pectin (by weight) after 30 days of storage at 25°C.

Basic **11.8** An online retailer needs to manage the amount of time needed to select the ordered items and assemble them for shipping. In order to assess the amount of time his assemblers devote to this task, the retailer takes a random sample of 100 orders and records the number of items in each order (NoItems) and the time needed to assemble the shipment.

NoItems	1	1	1	1	1	1	2	2	2	2	2	3	3	3	3
Time	4.8	14.7	13.7	9.5	2.4	11.8	10.5	12.0	14.3	16.4	17.9	14.5	16.7	20.1	16.8
NoItems	3	3	3	3	3	4	4	4	4	4	4	4	4	4	5
Time	20.8	26.8	12.9	13.0	15.4	10.8	20.3	22.3	20.0	21.9	20.7	20.9	19.9	17.1	19.1
NoItems	5	5	5	5	6	6	6	6	6	6	6	7	7	7	7
Time	17.2	18.1	17.9	12.9	19.6	27.9	22.2	23.2	17.5	15.3	21.8	17.9	20.1	18.9	29.3
NoItems	7	8	8	8	8	8	8	9	9	9	9	9	10	10	10
Time	21.5	26.5	28.2	25.1	26.1	28.2	22.3	21.8	25.5	24.4	18.1	26.7	24.6	30.1	21.5
NoItems	10	10	10	11	11	11	11	12	12	12	12	14	14	15	16
Time	25.2	28.2	22.6	31.3	27.2	28.8	29.7	34.3	27.9	29.7	28.7	34.6	38.0	28.0	37.0
NoItems	17	18	18	18	19	20	21	22	23	24	25	25	25	26	27
Time	32.5	39.5	39.0	37.0	35.5	38.3	44.0	39.6	42.3	34.6	44.9	47.4	49.2	45.8	44.0
NoItems	27	30	30	31	37	39	40	41	45	46					
Time	46.1	42.9	48.3	46.0	48.2	54.7	49.9	55.4	57.1	52.4					

a. Plot the data on a scatterplot.
b. Fit a least-squares line to the data, and comment on the degree of fit to the data.
c. Fit a regression model with the square root of NoItems as the explanatory variable.
d. Which model produced a better fit to the data?
e. Predict the amount of time needed to assemble a package containing 13 items using both models. Was there much difference in your predictions?

Engin. **11.9** A manufacturer of cases for sound equipment requires that holes be drilled for metal screws. The drill bits wear out and must be replaced; there is expense not only in the cost of the bits but also in the cost of lost production. Engineers varied the rotation speed of the drill and measured the lifetime y (thousands of holes drilled) of four bits at each of five speeds x. The data were:

x	60	60	60	60	80	80	80	80	100	100
y	4.6	3.8	4.9	4.5	4.7	5.8	5.5	5.4	5.0	4.5
x	100	100	120	120	120	120	140	140	140	140
y	3.2	4.8	4.1	4.5	4.0	3.8	3.6	3.0	3.5	3.4

a. Create a scatterplot of the data. Does there appear to be a relation? Does it appear to be linear?
b. Is there any evident outlier? If so, does it have high influence?

Engin. **11.10** Refer to Exercise 11.9.

```
Regression Analysis: Lifetime versus DrillSpeed

The regression equation is
Lifetime = 6.03 - 0.0170 DrillSpeed


Predictor        Coef    SE Coef       T      P
Constant       6.0300     0.5195   11.61  0.000
DrillSpeed  -0.017000   0.004999   -3.40  0.003


S = 0.632368   R-Sq = 39.1%   R-Sq(adj) = 35.7%


Analysis of Variance

Source          DF       SS       MS      F      P
Regression       1   4.6240   4.6240  11.56  0.003
Residual Error  18   7.1980   0.3999
Total           19  11.8220

Unusual Observations

Obs  DrillSpeed  LifeTime    Fit  SE Fit  Residual  St Resid
  2          60     3.800  5.010   0.245    -1.210    -2.08R

R denotes an observation with a large standardized residual.
```

a. Find the least-squares estimates of the slope and intercept in the output.
b. What does the sign of the slope indicate about the relation between the speed of the drill and bit lifetime?
c. Compute the residual standard deviation. What does this value indicate about the fitted regression line?

Engin. **11.11** Refer to the data of Exercise 11.9.

a. Use the regression line of Exercise 11.10 to calculate predicted values for $x = 60$, 80, 100, 120, and 140.
b. For which x-values are most of the actual y-values larger than the predicted y-values? For which x-values are most of the actual y-values smaller than the predicted y-values? What does this pattern indicate about whether there is a linear relation between the drill speed and the lifetime of the bit?
c. Suggest a transformation of the data to obtain a linear relation between the lifetime of the bit and the transformed values of the drill speed.

11.3 Inferences About Regression Parameters

Ag. **11.12** Refer to the data of Exercise 11.7.

a. Calculate a 95% confidence interval for β_1.
b. What is the interpretation of H_0: $\beta_1 = 0$ in Exercise 11.7?
c. Test the hypotheses H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$.
d. Determine the p-value of the test of H_0: $\beta_1 = 0$.

Ag. **11.13** Refer to the data of Exercise 11.7.

a. Calculate a 95% confidence interval for β_0.
b. What is the interpretation of H_0: $\beta_0 = 0$ for the problem in Exercise 11.7?
c. Test the hypotheses H_0: $\beta_0 = 0$ versus H_a: $\beta_0 \neq 0$.
d. Determine the p-value of the test of H_0: $\beta_0 = 0$.

Ag. **11.14** Refer to Exercise 11.7. Perform a statistical test of the null hypothesis that there is no linear relationship between the concentration of pectin and the firmness of canned sweet potatoes after 30 days of storage at 25°C. Give the p-value for this test and draw conclusions.

Bus. **11.15** Refer to the data of Exercise 11.8.

a. Calculate a 95% confidence interval for β_1.

b. What is the interpretation of H_0: $\beta_1 = 0$ in Exercise 11.8?

c. What is the natural research hypothesis H_a for the problem in Exercise 11.8?

d. Do the data support the research hypothesis from part (c) at $\alpha = .05$?

Bus. **11.16** Refer to the data of Exercise of 11.8.

a. Calculate a 95% confidence interval for β_0.

b. What is the interpretation of H_0: $\beta_0 = 0$ for the problem in Exercise 11.8?

c. Test the hypotheses H_0: $\beta_0 = 0$ versus H_a: $\beta_0 \neq 0$.

d. Determine the p-value of the test of H_0: $\beta_0 = 0$.

Bus. **11.17** Refer to Exercise 11.8. Perform a statistical test of the null hypothesis that there is no linear relationship between the time needed to select the ordered items and the number of items in the order. Give the p-value for this test, and draw conclusions.

Bio. **11.18** The extent of disease transmission can be affected greatly by the viability of infectious organisms suspended in the air. Because of the infectious nature of the disease under study, the viability of these organisms must be studied in an airtight chamber. One way to do this is to disperse an aerosol cloud, prepared from a solution containing the organisms, into the chamber. The biological recovery at any particular time is the percentage of the total number of organisms suspended in the aerosol that are viable. The data in the accompanying table are the biological recovery percentages computed from 13 different aerosol clouds. For each of the clouds, recovery percentages were determined at different times.

a. Plot the data.

b. Since there is some curvature, try to linearize the data using the log of the biological recovery.

Cloud	Time, x (in minutes)	Biological Recovery (%)
1	0	70.6
2	5	52.0
3	10	33.4
4	15	22.0
5	20	18.3
6	25	15.1
7	30	13.0
8	35	10.0
9	40	9.1
10	45	8.3
11	50	7.9
12	55	7.7
13	60	7.7

Bio. **11.19** Refer to Exercise 11.18.

a. Fit the linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, where y is the log biological recovery percentage.

b. Compute an estimate of σ_ε.

c. Identify the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Bio. **11.20** Refer to Exercise 11.18. Conduct a test of the null hypothesis that $\beta_1 = 0$. Use $\alpha = .05$.

Bio. **11.21** Refer to Exercise 11.18. Place a 95% confidence interval on β_0, the mean log biological recovery percentage at time zero. Interpret your findings. (*Note:* $E(y) = \beta_0$ when $x = 0$.)

Med. **11.22** Athletes are constantly seeking measures of the degree of their cardiovascular fitness prior to a major race. Athletes want to know when their training is at a level that will produce a peak performance. One such measure of fitness is the time to exhaustion from running on a

treadmill at a specified angle and speed. The important question is then "Does this measure of cardiovascular fitness translate into performance in a 10-km running race?" Twenty experienced distance runners who professed to be at top condition were evaluated on the treadmill and then had their times recorded in a 10-km race. The data are given here.

Treadmill time (minutes)	7.5	7.8	7.9	8.1	8.3	8.7	8.9	9.2	9.4	9.8
10-km time (minutes)	43.5	45.2	44.9	41.1	43.8	44.4	38.7	43.1	41.8	43.7
Treadmill time (minutes)	10.1	10.3	10.5	10.7	10.8	10.9	11.2	11.5	11.7	11.8
10-km time (minutes)	39.5	38.2	43.9	37.1	37.7	39.2	35.7	37.2	34.8	38.5

a. Plot the data in a scatterplot.
b. Fit a regression model to the data. Does a linear model seem appropriate?
c. Obtain the estimated linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$.

11.23 Refer to the data of Exercise 11.22.
a. Estimate σ_ε^2.
b. Estimate the standard error of $\hat{\beta}_1$.
c. Place a 95% confidence interval on β_1.
d. Test the hypothesis that there is a linear relationship between the amount of time needed to run a 10-km race and the time to exhaustion on a treadmill. Use $\alpha = .05$.

11.24 The focal point of an agricultural research study was the relationship between when a crop is planted and the amount of crop harvested. If a crop is planted too early or too late farmers may fail to obtain optimal yield and hence not make a profit. An ideal date for planting is set by the researchers, and the farmers then record the number of days either before or after the designated date. In the following data set, D is the deviation (in days) from the ideal planting date, and Y is the yield (in bushels per acre) of a wheat crop:

D	−11	−10	−9	−8	−7	−6	−4	−3	−1	0
Y	43.8	44.0	44.8	47.4	48.1	46.8	49.9	46.9	46.4	53.5
D	1	3	6	8	12	13	15	16	18	19
Y	55.0	46.9	44.1	50.2	41.0	42.8	36.5	35.8	32.2	33.3

a. Plot the above data. Does a linear relation appear to exist between yield and deviation from the ideal planting date?
b. Plot yield versus absolute deviation from the ideal planting date. Does a linear relation seem more appropriate in this plot than the plot in part (a)?

11.25 Refer to Exercise 11.24. Fit a regression model relating yield to the absolute deviation from the ideal planting date, that is, $x = |\text{D}|$.
a. Compute the estimated linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
b. Estimate σ_ε^2.
c. Estimate the standard error of $\hat{\beta}_1$.
d. Place a 95% confidence interval on β_1.
e. Test the hypothesis that there is a linear relationship between yield per acre and absolute deviation from the ideal planting date. Use $\alpha = .05$.

11.26 Refer to Exercise 11.24.
a. For this study, would it make sense to give any physical interpretation to β_0?
b. Place a 95% confidence interval on β_0, and give an interpretation to the interval relative to this particular study.
c. Test the hypotheses H_0: $\beta_0 = 0$ versus H_a: $\beta_0 \neq 0$. Does this test have any practical importance in this particular study?

Bus. **11.27** A firm that prints automobile bumper stickers conducts a study to investigate the relation between the direct cost of producing an order of bumper stickers (TOTCOST) and the number

of stickers (RunSize, in 1,000s of stickers) in a particular order. The data are given in the following table.

RunSize	2.6	5.0	10.0	2.0	.8	4.0	2.5	.6	0.8	1.0
TOTCOST	230	341	629	187	159	327	206	124	155	147
RunSize	2.0	3.0	.4	.5	5.0	20.0	5.0	2.0	1.0	1.5
TOTCOST	209	247	135	125	366	1,146	339	208	150	179
RunSize	.5	1.0	1.0	.6	2.0	1.5	3.0	6.5	2.2	1.0
TOTCOST	128	155	143	131	219	171	258	415	226	159

a. Plot a scatterplot of the data. Do you detect any difficulties with using a linear regression model? Can you find any blatant violations of the regression assumptions?
b. Compute the estimated regression line.
c. Estimate the residual standard deviation.
d. Construct a 95% confidence interval for the true slope.
e. What are the interpretations of the intercept and slope in this study?

11.28 Refer to Exercise 11.27.
a. Test the hypothesis H_0: $\beta_0 = 0$ using a t test with $\alpha = .05$.
b. Determine the p-value for this test, and interpret its value.

11.29 Refer to Exercise 11.27.
a. Compute the value of the F statistic and the associated p-value.
b. How do the p-values for this F statistic and the t test of Exercise 11.28 compare? Why should this relation hold?

11.4 Predicting New *y*-Values Using Regression

Bio. **11.30** Refer to Exercise 11.18. Using the least-squares line obtained in Exercise 11.18

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

estimate the mean log biological recovery percentage at 30 minutes using a 95% confidence interval.

Bio. **11.31** Use the data from Exercise 11.18 to complete the following.
a. Construct a 95% prediction interval for the log biological recovery percentage at 30 minutes.
b. Compare your results to the confidence interval on $E(y)$ from Exercise 11.30.
c. Explain the different interpretation for the two intervals.

Engin. **11.32** A chemist is interested in determining the weight loss y of a particular compound as a function of the amount of time the compound is exposed to the air. The data in the following table give the weight losses associated with $n = 12$ settings of the independent variable, exposure time.

Weight Loss and Exposure Time Data

Weight Loss, *y* (in pounds)	Exposure Time (in hours)	Weight Loss, *y* (in pounds)	Exposure Time (in hours)
4.3	4	6.6	6
5.5	5	7.5	7
6.8	6	2.0	4
8.0	7	4.0	5
4.0	4	5.7	6
5.2	5	6.5	7

a. Determine the least-squares prediction equation for the model

$y = \beta_0 + \beta_1 x + \varepsilon$.

b. Test H_0: $\beta_1 \leq 0$; give the p-value for H_a: $\beta_1 > 0$, and draw conclusions.

Engin. **11.33** Refer to Exercise 11.32.

a. Determine the 95% confidence bands for $E(y)$ when $4 \leq x \leq 7$.

b. Determine the 95% prediction bands for y, $4 \leq x \leq 7$.

c. Distinguish between the meaning of the confidence bands and the prediction bands in parts (a) and (b).

Engin. **11.34** Refer to Exercise 11.27.

a. Predict the mean total direct cost for all bumper sticker orders with a print run of 2,000 stickers (that is, with RunSize = 2.0).

b. Compute a 95% confidence interval for this mean.

Engin. **11.35** Refer to Exercise 11.27.

a. Predict the direct cost for a particular bumper sticker order with a print run of 2,000 stickers. Obtain a 95% prediction interval.

b. Would an actual direct cost of $250 be surprising for this order?

Med. **11.36** Use the data from Exercise 11.22.

a. Estimate the mean time to run 10 km for athletes having a treadmill time of 11 minutes.

b. Place a 95% confidence interval on the mean time to run 10 km for athletes having a treadmill time of 11 minutes.

Med. **11.37** Refer to Exercise 11.22 to complete the following.

a. Predict the time to run 10 km if an athlete has a treadmill time of 11 minutes.

b. Place a 95% prediction interval on the time to run 10 km for an athlete having a treadmill time of 11 minutes.

c. Compare the 95% prediction interval from part (b) to the 95% confidence interval from Exercise 11.36. What is the difference in the interpretation of these two intervals? Provide a nontechnical reason why the prediction interval is wider than the confidence interval.

11.5 Examining Lack of Fit in Linear Regression

Engin. **11.38** A manufacturer of laundry detergent was interested in testing a new product prior to market release. One area of concern was the relationship between the height of the detergent suds in a washing machine as a function of the amount of detergent added in the wash cycle. For a standard-size washing machine tub filled to the full level, the manufacturer made random assignments of amounts of detergent and tested them on the washing machine. The data appear next.

Height, y	Amount, x
28.1, 27.6	6
32.3, 33.2	7
34.8, 35.0	8
38.2, 39.4	9
43.5, 46.8	10

a. Plot the data.

b. Fit a linear regression model.

c. Use a residual plot to investigate possible lack of fit.

11.39 Refer to Exercise 11.38.

a. Conduct a test for lack of fit of the linear regression model.

b. If the model is appropriate, give a 95% prediction band for y.

11.40 In the preliminary studies of a new drug, a pharmaceutical firm needs to obtain information on the relationship between the dose level and potency of the drug. In order to obtain this information, a total of 18 test tubes are inoculated with a virus culture and incubated for an appropriate period of time. Three test tubes are randomly assigned to each of six different dose levels. The 18 test tubes are then injected with the randomly assigned dose level of the drug. The measured response is the protective strength of the drug against the virus culture. Due to a problem with a few of the test tubes, only two responses were obtained for dose levels 4, 8, and 32. The data are given here:

Dose level	2	4	8	16	32	64
Response	5, 7, 3	10, 14	15, 17	20, 21, 19	23, 29	28, 31, 30

a. Plot the data.
b. Fit a linear regression model to these data.
c. From a plot of the residuals, does there appear to be a possible lack of fit of the linear model?

11.41 Refer to Exercise 11.40. Conduct a test for lack of fit of the linear regression model.

11.42 Refer to Exercise 11.40. Often in drug evaluations, a logarithmic transformation of the dose levels will yield a linear relationship between the response variable and the independent variable. Let x_i be the natural logarithm of the dose levels, and evaluate the regression of the response of the drug in the 15 test tubes to the transformed independent variable: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

a. Plot the response of the drug versus the natural logarithm of the dose levels. Does it appear that a linear model is appropriate?
b. Fit a linear regression model to these data.
c. From a plot of the residuals, do these appear to be a possible lack of fit of the linear model?
d. Conduct a test for lack of fit of the linear regression model.

11.6 Correlation

11.43 Refer to Exercise 11.27.

a. Compute the value of r^2_{yx}.
b. What are the value and sign of the correlation coefficient?
c. Suppose the study in Exercise 11.27 had been restricted to RunSize values less than 1.8. Would you anticipate a larger or smaller value for the correlation coefficient? Explain your answer.

Edu.

11.44 A survey of MBA. graduates of a business school obtained data on the first-year salary after graduation and years of work experience prior to obtaining their MBA. The data are given in the following table with salary in thousands of dollars.

EXPER	8	5	5	11	4	3	3	3	0	13	14	10	2
SALARY	113.9	112.5	109	125.1	111.6	112.7	104.5	100.1	101.1	126.9	97.9	113.5	98.3
EXPER	2	5	13	1	5	1	5	5	7	4	3	3	7
SALARY	97.2	111.3	124.7	105.3	107	103.8	107.4	100.2	112.8	100.7	107.3	103.7	121.8
EXPER	7	9	6	6	4	6	5	1	13	1	6	2	4
SALARY	111.7	116.2	108.9	111.9	96.1	113.5	110.4	98.7	120.1	98.9	108.4	110.6	101.8
EXPER	1	5	1	4	1	2	7	5	1	1	0	1	6
SALARY	104.4	106.6	103.9	105	97.9	104.6	106.9	107.6	103.2	101.6	99.2	101.7	120.1

a. Plot the data in a scatterplot. Based on the plotted data, does it appear that those students having less experience also have smaller salaries?
b. Identify any students who do not seem to satisfy the pattern of larger salaries associated with more experience.

Edu. **11.45** Refer to the data in Exercise 11.44.
a. Compute the correlation coefficient between years of experience and first-year salary. Do the sign and size of the correlation agree with the pattern observed in the scatterplot?
b. Compute the Spearman rank correlation coefficient between years of experience and first-year salary.
c. Which of the correlations is more influenced by the data values that do not follow the overall pattern?

Edu. **11.46** Refer to the data in Exercise 11.44.
a. Determine the least-squares estimates of the slope and intercept in the regression line relating first-year salary to years of experience. Interpret the coefficients. Is the intercept meaningful in the context of this data set?
b. Compute the residual standard deviation. Interpret this value.
c. Is there a significant relationship between salary and experience?
d. How much of the variability in salaries is accounted for by the number of years of experience?

Edu. **11.47** Refer to the data in Exercise 11.44. The student with 14 years of experience with a starting salary of $97,900 was hired by a family business. In return for a low starting salary, the student received a large equity share in the firm.
a. Would the data value associated with this student be considered a high leverage or a high influence data value?
b. Would the slope increase or decrease if this point was removed from the analysis?
c. In which direction (larger or smaller) would the removal of this data point change the residual standard deviation?
d. How would the removal of this data point change the correlation?

Edu. **11.48** Refer to the data in Exercise 11.47.
a. Refit the regression model with the data value (14, 97.9) removed. How large were the changes in the slope and residual standard deviation?
b. Compute the correlation coefficient with the data value (14, 97.9) removed. How large was the change in the correlation coefficient compared with the value computed from all the data?
c. Compute the Spearman rank correlation coefficient for the complete data set and for the data set with the value (14, 97.9) removed.
d. Was the change in the Spearman rank correlation coefficient larger or smaller than the change in the standard correlation coefficient?

11.49 Refer to Example 6.7. In this example, an insurance adjuster wanted to know the degree to which the two garages were in agreement on their estimates of automobile repairs. The data given below are the estimated costs from the two garages for repairing 15 cars.

Car	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Garage I	17.6	20.2	19.5	11.3	13.0	16.3	15.3	16.2	12.2	14.8	21.3	22.1	16.9	17.6	18.4
Garage II	17.3	19.1	18.4	11.5	12.7	15.8	14.9	15.3	12.0	14.2	21.0	21.0	16.1	16.7	17.5

a. Compute the correlation between the car repair estimates from the two garages.
b. Calculate a 95% confidence interval for the correlation coefficient.
c. Does the very large positive value for the correlation coefficient indicate that the two garages are providing nearly identical estimates for the repairs? If not, explain why this statement is wrong.

Edu. **11.50** There has been an increasing emphasis in recent years on making sure that young women are given the same opportunities to develop their mathematical skills as young men are given in U.S. educational systems. The following table provides the SAT scores for male and female students over a 34-year period.

Gender/Type	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Male/Verbal	506	508	509	508	511	514	515	512	512	510
Female/Verbal	498	496	499	498	498	503	504	502	499	498
Male/Math	515	516	516	516	518	522	523	523	521	523
Female/Math	473	473	473	474	478	480	479	481	483	482
Gender/Type	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
Male/Verbal	505	503	504	504	501	505	507	507	509	509
Female/Verbal	496	495	496	497	497	502	503	503	502	502
Male/Math	521	520	521	524	523	525	527	530	531	531
Female/Math	483	482	484	484	487	490	492	494	496	495
Gender/Type	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
Male/Verbal	507	509	507	512	512	513	505	503	502	502
Female/Verbal	504	502	502	503	504	505	502	500	499	497
Male/Math	533	533	534	537	537	538	536	532	532	533
Female/Math	498	498	500	503	501	504	502	499	499	498
Gender/Type	**2010**	**2011**	**2012**	**2013**						
Male/Verbal	502	500	498	499						
Female/Verbal	498	495	493	494						
Male/Math	533	531	532	531						
Female/Math	499	500	499	499						

Source: ***CollegeBoard. (2013).* Total Group Profile Report.**

a. Plot the six pairs of data values in scatterplots: Male/Verbal versus Female/Verbal, Male/Math versus Male/Verbal, and so on.
b. Which , if any, of the six correlations are significantly different from 0 at the 5% level?
c. Do the plots reflect the sizes of the correlations between the pairs of variables?
d. Are male verbal scores more correlated with male or female math scores?

Edu. **11.51** Refer to Exercise 11.50.

a. Place a 95% confidence interval on the six correlations.
b. Using the confidence intervals from part (b), are there any differences in the degree of correlation between male and female math scores?
c. Using the confidence intervals from part (b), are there any differences in the degree of correlation between male and female verbal scores?
d. Are your answers to parts (b) and (c) different from your answer to part (c) in Exercise 11.50?

Supplementary Exercises

11.52 A construction science class project was to compare the daily gas consumption of 20 homes with a new form of insulation to that of 20 similar homes with standard insulation. The students set up instruments to record the temperature both inside and outside of the homes over a 6-month period of time (October–March). The average differences in these values are given below. The students also obtained the average daily gas consumption (in kilowatt hours). All the homes were heated with gas. The data are given here:

Data for Homes with Standard Form of Insulation:

TempDiff (°F)	20.3	20.7	20.9	22.8	23.1	24.8	25.9	26.1	27.0	27.2
GasConsumption (kWh)	70.3	70.7	72.9	77.6	79.3	86.5	90.6	91.9	94.5	92.7
TempDiff (°F)	29.8	30.2	30.6	31.8	33.2	33.4	34.2	35.1	36.2	36.5
GasConsumption (kWh)	104.8	103.2	91.2	89.6	116.2	116.9	105.1	106.1	117.8	120.3

Data for Homes with New Form of Insulation:

TempDiff (°F)	20.1	21.1	21.9	22.6	23.4	24.2	24.9	25.1	26.0	27.2
GasConsumption (kWh)	65.3	66.5	67.8	73.2	75.3	81.1	82.2	85.7	90.9	87.4
TempDiff (°F)	28.8	29.2	30.6	30.8	32.6	32.4	34.8	35.9	36.0	36.5
GasConsumption (kWh)	94.9	93.9	87.1	84.2	106.6	111.3	100.9	101.9	110.1	119.1

a. Obtain the estimated regression lines for the two types of insulation.
b. Compare the fits of the two lines.
c. Is the rate of increase in gas consumption as temperature difference increases less for the new type of insulation? Justify your answer by using 95% confidence intervals.
d. If the rates are comparable, describe how the two lines differ.

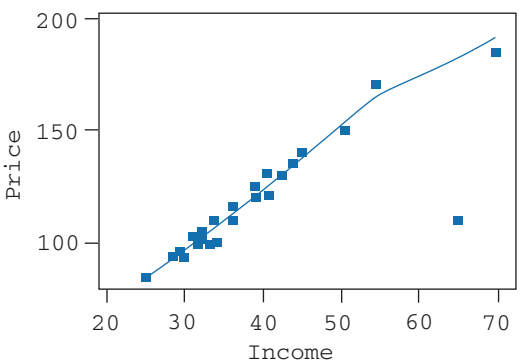
11.53 Refer to Exercise 11.52.
a. Predict the average gas consumption for both groups of homes when the temperature difference is 20°F.
b. Place 95% confidence intervals on your predicted values in part (a).
c. Based on the two confidence intervals, do you believe that the average gas consumption has been reduced by using the new form of insulation?
d. Predict the gas consumption of a home insulated with the new type of insulation if the temperature difference was 50°F.

Bio.

11.54 A realtor studied the relation between x = yearly income (in thousands of dollars per year) of home purchasers and y = sale price of the house (in thousands of dollars). The realtor gathered data from mortgage applications for 24 sales in the realtor's basic sales area in one season.

x	25.0	28.5	29.2	30.0	31.0	31.5	31.9	32.0	33.0
y	84.9	94.0	96.5	93.5	102.9	99.5	101.0	105.0	99.9
x	33.5	34.0	35.9	36.0	39.0	39.0	40.5	40.9	42.5
y	110.0	100.0	116.0	110.0	125.0	119.9	130.6	120.8	129.9
x	44.0	45.0	50.0	54.6	65.0	70.0			
y	135.5	140.0	150.7	170.0	110.0	185.0			

a. A scatterplot with a LOWESS smoother, drawn using Minitab, follows. Does the relation appear to be basically linear?
b. Are there any high leverage points? If so, which ones seem to have high influence?



11.55 For Exercise 11.54,

a. Determine the least-squares regression equation for the data.

b. Interpret the slope coefficient. Is the intercept meaningful?

c. Compute the residual standard deviation.

Edu. **11.56** Refer to Exercise 11.54. Delete the data value $x = 65.0$, $y = 110.0$ from the data set.

a. Refit the regression line, and compare the slope of the line with and without the data value $x = 65.0$, $y = 110.0$ in the set.

b. Compute the two forms of the correlation coefficient, and compare their values.

c. Is the Spearman rank correlation coefficient less or more affected by an extreme value compared to the standard correlation coefficient?

Ag. **11.57** A researcher conducts an experiment to examine the relationship between the weight gain of chickens whose diets had been supplemented by different amounts of the amino acid lysine and the amount of lysine ingested. Since the percentage of lysine is known and we can monitor the amount of feed consumed, we can determine the amount of lysine eaten. A random sample of 12 2-week-old chickens was selected for the study. Each was caged separately and was allowed to eat at will from feed composed of a base supplemented with lysine. The sample data summarizing weight gains and amounts of lysine eaten over the test period are given here. (In the data, y represents weight gain in grams, and x represents the amount of lysine ingested in grams.)

a. From the scatterplot of the data, does a linear model seem appropriate?

b. Compute the estimated linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Chick	Weight Gain, y (in grams)	Lysine Ingested, x (in grams)	Chick	Weight Gain, y (in grams)	Lysine Ingested, x (in grams)
1	14.7	.09	7	17.2	.11
2	17.8	.14	8	18.7	.19
3	19.6	.18	9	20.2	.23
4	18.4	.15	10	16.0	.13
5	20.5	.16	11	17.8	.17
6	21.1	.23	12	19.4	.21

11.58 Refer to Exercise 11.57.

a. Estimate σ_ε^2.

b. Compute the standard error of $\hat{\beta}_1$.

c. Conduct a statistical test of the research hypothesis that for this diet preparation and length of study, there is a direct (positive) linear relationship between weight gain and amount of lysine eaten.

11.59 Refer to Exercise 11.57.

a. For this exercise, would it make sense to give any physical interpretation to β_0? (*Hint:* The lysine was mixed in the feed.)

b. Consider an alternative model relating weight gain to amount of lysine ingested:

$$y = \beta_1 x + \varepsilon$$

Distinguish between this model and the model $y = \beta_0 + \beta_1 x + \varepsilon$.

11.60 **a.** Refer to part (b) of Exercise 11.59. Obtain $\hat{\beta}_1$ for the model $y = \beta_1 x + \varepsilon$, where

$$\hat{\beta}_1 = \frac{\sum xy}{\sum x^2}$$

b. Which of the two models, $y = \beta_0 + \beta_1 x + \varepsilon$ or $y = \beta_1 x + \varepsilon$, appears to give a better fit to the sample data? (*Hint:* Examine the two prediction equations on a graph of the sample observations.)

Engin. **11.61** An air conditioning company responds to calls concerning problems with air conditioners by sending a repair person to the home of the caller. There have been complaints about lengthy delays between the time the call is received and the time when the repair person reports to the home.

The manager of the company would like to develop a method to estimate the length of time the customer will have to wait before receiving service. Data is obtained by taking a random sample of 15 calls for service for each backlog situation in which 0, 1, 2, 3, or 4 previous callers are waiting for service and then recording the number of minutes it took for the service person to reach the customer.

Backlog	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Resp. time	2	3	5	6	8	10	13	15	16	20	24	27	32	35	42
Backlog	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Resp. time	12	23	51	36	48	112	123	163	172	120	252	237	212	245	246
Backlog	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Resp. time	42	38	105	156	158	210	183	215	216	320	324	278	332	375	412
Backlog	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Resp. time	62	73	58	126	208	270	313	415	416	320	324	427	432	435	442
Backlog	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Resp. time	82	93	105	206	278	310	313	415	316	420	424	527	532	635	642

a. Plot the data, and assess whether fitting a regression model relating response time to backlog would be appropriate.
b. Fit a regression line relating the response time to the backlog of previous calls.
c. Fit a regression line relating the logarithm of response time to the backlog of previous calls.
d. Which of the two regression lines appears to be most appropriate?

Engin. **11.62** Refer to Exercise 11.61.
a. Calculate the predicted response time if there is a backlog of four customers.
b. Place a 95% prediction interval on your prediction in part (a).
c. Compute a 95% confidence interval on the mean response time for situations where there is a backlog of four customers. Compare this interval to the interval computed in part (b).
d. What is the difference in interpretation of the two intervals computed in parts (b) and (c)?
e. The manager has requested an estimate of the mean response time if there was a backlog of seven customers. What is the problem with producing the estimate?

Engin. **11.63** Refer to Exercise 11.61.
a. Test for lack of fit for the model relating response time to backlog.
b. Test for lack of fit for the model relating logarithm of response time to backlog.
c. Are the results from parts (a) and (b) consistent with the patterns observed in the scatterplots?

Engin. **11.64** Refer to Exercise 11.61.
a. Compute the standard correlation coefficient, r_{yx}, between the backlog and response time.
b. Compute the standard correlation coefficient, r_{yx}, between the backlog and logarithm of response time.
c. Compute the Spearman rank correlation coefficient, r_s, between the backlog and response time.
d. Compute the Spearman rank correlation coefficient, r_s, between the backlog and logarithm of response time.
e. Which of the two correlations best reflects the relationship between the backlog and response time?

Env. **11.65** An airline designs a study to evaluate fuel usage by a certain type of aircraft. From a random sample of 50 flights, the flight length in hundreds of miles and the fuel usage in gallons are recorded.

Mileage	530	533	536	569	580	603	655	667	707	712
FuelUse	382	257	376	290	416	362	361	347	498	449
Mileage	735	784	814	839	844	885	890	913	957	976
FuelUse	482	452	426	441	524	488	551	570	556	522
Mileage	979	1,050	1,055	1,069	1,070	1,114	1,116	1,129	1,308	1,348
FuelUse	542	640	598	502	639	679	630	659	695	767
Mileage	1,356	1,363	1,395	1,474	1,504	1,528	1,613	1,615	1,632	1,657
FuelUse	632	641	740	737	783	802	861	874	847	748
Mileage	1,674	1,698	1,730	1,769	1,775	1,789	1,804	1,820	1,851	1,983
FuelUse	872	802	925	912	936	846	883	902	925	908

a. Plot fuel usage versus and mileage. Does the plot display a linear relationship between fuel usage and length of flight?
b. Obtain a regression equation relating fuel usage to length of the flight.
c. What is the interpretation of $\hat{\beta}_1$ in this situation?
d. Is there a sensible interpretation of $\hat{\beta}_0$ in this situation?
e. Compute the correlation coefficient, r_{yx}, and the coefficient of determination. Interpret these values.

Env. **11.66** Refer to Exercise 11.65.

a. Estimate the mean fuel usage for a 1000-mile flight. Provide a 95% confidence interval for your estimate.
b. Predict the fuel usage for a particular 1000-mile flight. Would a fuel usage of 700 gallons be considered excessive?
c. The airline is considering a new flight from New York to Paris. Provide a prediction of the amount of fuel to be used in this flight. The flying distance from New York to Paris is 3,500 miles.

Env. **11.67** Refer to Exercise 11.65.

a. What are some of the other variables that would be related to fuel usage that may improve the fit of the regression line?
b. How could you measure the improvement in the fit of the regression model?

Ag. **11.68** A forester has a unique ability to estimate the volume (in cubic feet) of trees prior to a timber sale. The timber company that employs the forester wants him to train other employees in his technique of estimation. After a training period, the forester randomly selects 25 trees that will be cut down for processing. The forester's assistant estimates the cubic-foot volume of each tree. After the tree has been chopped down, the forester obtains its actual cubic-foot volume.

Estimated volume	11.1	13.0	12.0	11.2	12.2	13.0	12.5	16.2	14.4	15.4	15.9	16.4	15.5
Actual volume	11.4	12.5	13.1	13.3	13.7	13.8	14.3	15.9	16.4	17.0	18.8	18.8	19.2
Estimated volume	16.9	17.6	15.8	16.4	18.7	18.9	19.7	19.7	21.0	19.0	21.5	21.3	
Actual volume	19.7	19.8	19.8	20.1	20.1	20.9	22.4	22.7	23.1	23.3	24.0	24.8	

a. Plot the data in a scatterplot. Does there appear to be a reasonable relation between the estimated and actual volumes?
b. Fit a regression model relating the estimated volume to the actual volume.
c. If the assistant is producing very accurate estimates of the volume, what should be the value of the slope of the regression line?
d. Is there significant evidence that the assistant is producing accurate estimates of the volume of the trees?

Ag. **11.69** Refer to Exercise 11.68.

a. Predict the actual cubic-foot volume for a tree that the assistant estimates to have a cubic-foot volume of 13?

b. Place a 95% prediction interval on the actual cubic-foot volume for a tree that the assistant estimate to have a cubic-foot volume of 13.

Med. **11.70** A research MD designs a study to examine the relationship between the dose of a drug and the cumulative urine volume (CUMVOL) for a drug being considered as a diuretic. The selected group of 24 patients yields the following results.

Dose	6	6	6	6	6	6	9	9	9	9	9	9
CUMVOL	7.1	11.5	8.4	8.0	9.4	12.0	13.2	14.7	12.7	15.5	18.4	14.4
Dose	13.5	13.5	13.5	13.5	13.5	13.5	20.25	20.25	20.25	20.25	20.25	20.25
CUMVOL	12.1	15.8	13.8	20.4	22.7	17.0	19.8	15.6	25.3	13.5	24.8	20.9

a. Plot the data in a scatterplot. Would a straight line be an appropriate model relating dose to CUMVOL?

b. Fit a regression model relating CUMVOL to dose.

c. Test for lack of fit of the model at the $\alpha = .05$ level.

d. Estimate the mean value of CUMVOL for a dose level of 15 using a 95% confidence interval

Med. **11.71** Refer to Exercise 11.70.

a. The researcher consulted with a statistician, and a transformation of the data was suggested. Plot the square root of CUMVOL versus the logarithm of dose in a scatterplot. Do the plotted points appear to be more closely related by a straight line than were the raw data values?

b. Fit a regression model relating the square root of CUMVOL to the logarithm of dose.

c. Test for lack of fit of this model at the $\alpha = .05$ level.

d. Estimate the mean value of CUMVOL for a dose level of 15 using a 95% confidence interval based on the model obtained in part (b).

e. How large are the differences in the two estimates of the mean CUMVOL?

Med. **11.72** Refer to Exercise 11.70.

a. Estimate the dose level needed to produce a CUMVOL of 20.

b. Place a 95% confidence interval on your estimate.

Engin. **11.73** The management science staff of a grocery products manufacturer is developing a linear programming model for the production and distribution of its cereal products. The model requires transportation costs for a very large number of origins and destinations. It is impractical to do the detailed tariff analysis for every possible combination, so a sample of 48 routes is selected. For each route, the mileage x and shipping rate y (in dollars per 100 pounds) are found.

The data are as follows:

Mileage	50	60	80	80	90	90	100	100	100	110	110	110
Rate	12.7	13.0	13.7	14.1	14.6	14.1	15.6	14.9	14.5	15.3	15.5	15.9
Mileage	120	120	120	120	130	130	140	150	170	190	200	230
Rate	16.4	11.1	16.0	15.8	16.0	16.7	17.2	17.5	18.6	19.3	20.4	21.8
Mileage	260	300	330	340	370	400	440	440	480	510	540	600
Rate	24.7	24.7	18.0	27.1	28.2	30.6	31.8	32.4	34.5	35.0	36.3	41.4
Mileage	650	700	720	760	800	810	850	920	960	1,050	1,200	1,650
Rate	46.4	45.8	46.6	48.0	51.7	50.2	53.6	57.9	56.1	58.7	75.8	89.0

a. Obtain the regression equation and the residual standard deviation.

b. Calculate a 90% confidence interval for the true slope.

11.74 In a scatterplot of the data from Exercise 11.73, do you see any problems with the data?

11.75 For Exercise 11.73, predict the shipping rate for a 350-mile route. Obtain a 95% prediction interval. How serious is the extrapolation problem in this exercise?

Soc.

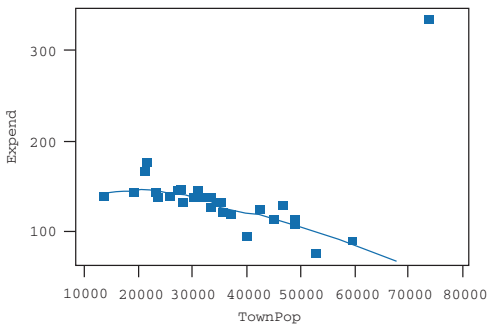
11.76 Suburban towns often spend a large fraction of their municipal budgets on public safety (police, fire, and ambulance) services. A taxpayers' group felt that very small towns were likely to spend large amounts per person because they have such small financial bases. The group obtained data on the per capita expenditure for public safety of 29 suburban towns in a metropolitan area, as well as the population of each town in units of 10,000 people.

TownPop	14	20	22	22	24	24	26	28	29	30
Expend	140	142	165	175	143	141	142	144	144.5	138
TownPop	30	31	32	32	32	32	34	34	36	36
Expend	139	141	140	139	137	137.2	137.0	136.5	136	135.5
TownPop	38	40	43	45	49	49.5	52	60	76	
Expend	105	132	128	135	129	126	70	95	310	

a. If the taxpayers' group is correct, what sign should the slope of the regression model have?

b. Does the slope in the output confirm the opinion of the group?

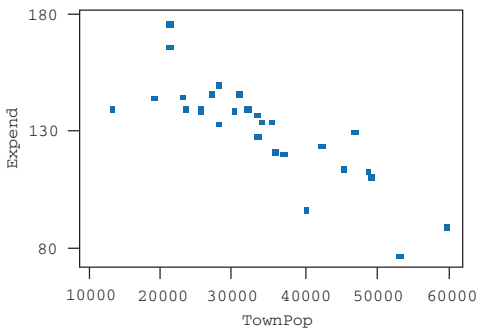
11.77 Minitab produced a scatterplot and LOWESS smoothing of the data in Exercise 11.76, shown here. Does this plot indicate that the regression line is misleading? Why?



11.78 One town in the data base of Exercise 11.76 is the home of an enormous regional shopping mall. A very large fraction of the town's expenditure on public safety is related to the mall; the mall management pays a yearly fee to the township that covers these expenditures. That town's data were removed from the data base and the remaining data were reanalyzed by Minitab. A scatterplot is shown.

a. Explain why removing this one point from the data changed the regression line so substantially.

b. Does the revised regression line appear to conform to the opinion of the taxpayers' group in Exercise 11.76?



Soc. **11.79** Refer to Exercise 11.76.

a. Obtain the regression line with the one unusual town removed from the data set.

b. Estimate the expenditure on public safety for a town of 37,000 people. Compare this estimate with an estimate using the complete data set.

c. Compare the estimated slope from the regression fit using the data set with the unusual town removed to the estimated slope from the regression fit using the complete data set? Discuss the impact of an extreme data value on the reliability of the inferences that can be made from the data about the population from which the data were obtained.

Bio. **11.80** In screening for compounds useful in treating hypertension (high blood pressure), researchers assign six rats to each of three groups. The rats in group 1 receive .1 mg/kg of a test compound; those in groups 2 and 3 receive .2 and .4 mg/kg, respectively. The response of interest is the decrease in blood pressure 2 hours postdose compared to the corresponding predose blood pressure. The data are shown here:

	Dose, x	Blood Pressure Drop, y (in mm Hg)					
Group 1	.1 mg/kg	10	12	15	16	13	11
Group 2	.2 mg/kg	25	22	26	19	18	24
Group 3	.4 mg/kg	30	32	35	27	26	29

a. Fit the following model to the data.

$$y = \beta_0 + \beta_1 \log_{10} y + \varepsilon$$

b. Use residual plots to examine the fit to the model in part (a).

c. Conduct a statistical test of H_0: $\beta_1 \leq 0$ versus H_a: $\beta_1 > 0$. Give the p-value for your test.

Ag. **11.81** A laboratory conducts a study to examine the effect of different levels of nitrogen on the yield of lettuce plants. Use the data shown here to fit a linear regression model. Test for possible lack of fit of the model.

Coded Nitrogen	Yield (Emergent Stalks per Plot)
1	21, 18, 17
2	24, 22, 26
3	34, 29, 32

Med. **11.82** Researchers measured the specific activity of the enzyme sucrase extracted from portions of the intestines of 24 patients who underwent an intestinal bypass. After the sections were extracted, they were homogenized and analyzed for enzyme activity (Carter, 1981). Two different methods can be used to measure the activity of sucrase: the homogenate method and the pellet method. Data for the 24 patients are shown here for the two methods:

Sucrase Activity as Measured by the Homogenate and Pellet Methods

Patient	Homogenate Method, y	Pellet Method, x
1	18.88	70.00
2	7.26	55.43
3	6.50	18.87
4	9.83	40.41
5	46.05	57.43
6	20.10	31.14
7	35.78	70.10
8	59.42	137.56
9	58.43	221.20
10	62.32	276.43

(continued)

Sucrase Activity as Measured by the Homogenate and Pellet Methods

Patient	Homogenate Method, y	Pellet Method, x
11	88.53	316.00
12	19.50	75.56
13	60.78	277.30
14	77.92	331.50
15	51.29	133.74
16	77.91	221.50
17	36.65	132.93
18	31.17	85.38
19	66.09	142.34
20	115.15	294.63
21	95.88	262.52
22	64.61	183.56
23	37.71	86.12
24	100.82	226.55

a. Produce a scatterplot of the data. Might a linear model adequately describe the relationship between the two methods?
b. Produce a residual plot. Are there any potential problems uncovered by the plot?
c. In general, the pellet method is more time consuming than the homogenate method, yet it provides a more accurate measure of sucrase activity. How might you estimate the pellet reading based on a particular homogenate reading?
d. How would you develop a confidence (prediction) interval about your point estimate?

Bus. **11.83** A realtor in a suburban area attempted to predict house prices solely on the basis of size. From a listing service, the realtor obtained size in thousands of square feet and asking price in thousands of dollars.

Price	210	145	168	352	234	148	217	216	213	143	178	131	181	148	127	158	226	194	166
Size	2.5	1.5	1.8	4.7	2.4	1.5	2.5	3.3	2.6	1.6	1.6	1.4	2.9	1.6	1.9	1.7	2.6	1.9	1.8
Price	207	139	143	141	142	214	262	191	167	153	153	184	123	182	143	144	161	157	155
Size	2.8	1.5	1.5	1.9	1.6	2.2	2.7	2.0	2.2	1.6	1.6	2.3	1.4	1.9	1.6	1.5	1.6	1.7	1.7
Price	203	147	173	160	219	156	169	133	154	220	151	188	153	215	144	125	152	132	164
Size	2.2	1.8	1.8	1.7	2.4	1.9	1.9	1.5	2.9	2.9	1.9	2.3	1.7	2.1	1.9	1.7	1.7	1.4	2.0

a. Obtain a plot of price against size. Does it appear there is an increasing relation?
b. Locate an apparent outlier in the data. Is it a high leverage point?
c. Obtain a regression equation, and include the outlier in the data.
d. Delete the outlier, and obtain a new regression equation. How much does the slope change without the outlier? Why?
e. Locate the residual standard deviations for the outlier-included and outlier-excluded models. Do they differ much? Why?

11.84 Obtain the outlier-excluded regression model for the data of Exercise 11.83.
a. Interpret the intercept (constant) term. How much meaning does this number have in this context?
b. What would it mean in this context if the slope was 0? Can the null hypothesis of zero slope be emphatically rejected?
c. Calculate a 95% confidence interval for the true population value of the slope.

11.85 **a.** Obtain a 95% prediction interval for the asking price of a home of 5,000 square feet, based on the outlier-excluded data of Exercise 11.83. Would this be a wise prediction to make, based on the data?
b. Obtain a plot of the price against the size. Does the constant-variance assumption seem reasonable, or does variability increase as size increases?
c. What does your answer to part (b) say about the prediction interval obtained in part (a)?

Bus. **11.86** A lawn care company tried to predict the demand for its service by zip code, using the housing density in the zip code area as a predictor. The owners obtained the number of houses and the geographic size of each zip code and calculated their sales per thousand homes and number of homes per acre.

Sales	54	72	54	62	72	83	115	90	66	60	100	78	152	87	54	82
Density	6.5	4.6	5.5	4.6	4.2	4.3	2.3	3.5	3.2	8.4	3.4	4.0	2.0	3.2	6.7	3.0
Sales	59	183	171	96	134	79	94	82	66	62	45	69	65	81	94	117
Density	5.7	1.3	1.3	3.0	2.2	4.3	2.6	3.0	4.3	7.8	9.4	4.2	5.9	6.2	2.8	2.4

a. Obtain the correlation between the two variables. What does its sign mean?
b. Obtain a prediction equation with sales as the dependent variable and density as the independent variable. Interpret the intercept (yes, we know the interpretation will be a bit strange) and the slope numbers.
c. Obtain a value for the residual standard deviation. What does this number indicate about the accuracy of prediction?

11.87 **a.** Obtain a value of the t statistic for the regression model of Exercise 11.86. Is there conclusive evidence that density is a predictor of sales?
b. Calculate a 95% confidence interval for the true value of the slope. The package should have calculated the standard error for you.

11.88 Obtain a plot of the data of Exercise 11.86 with sales plotted against density. Does it appear that straight-line prediction makes sense?

11.89 Refer to Exercise 11.86. Calculate a new variable: $x = 1/\text{density}$.
a. What is the interpretation of the new variable? In particular, if the new variable equals 0.50, what does that mean about the particular zip code area?
b. Plot sales against the new variable. Does a straight-line prediction look reasonable here?
c. Obtain the correlation of sales and the new variable. Compare its magnitude to the correlation obtained in Exercise 11.86 between sales and density. What explains the difference?

Engin. **11.90** A manufacturer of paint used for marking road surfaces developed a new formulation that needs to be tested for durability. One question concerns the concentration of pigment in the paint. If the concentration is too low, the paint will fade quickly; if the concentration is too high, the paint will not adhere well to the road surface. The manufacturer applies paint at various concentrations to sample road surfaces and obtains a durability measurement for each sample.

Conc.	20	20	20	20	20	20	20	20	20	20	20	20
Durab.	53.3	25.2	41.9	20.3	55.5	50.7	57.1	34.1	52.7	42.5	51.7	47.0
Conc.	30	30	30	30	30	30	30	30	30	30	30	30
Durab.	67.2	66.7	56.7	60.3	68.0	56.1	59.9	63.3	64.4	49.3	61.7	62.3
Conc.	40	40	40	40	40	40	40	40	40	40	40	40
Durab.	64.7	68.0	76.5	69.9	69.1	50.7	57.1	65.7	67.1	74.4	73.5	69.9
Conc.	50	50	50	50	50	50	50	50	50	50	50	50
Durab.	51.6	75.7	55.9	76.1	55.3	73.3	61.5	53.3	74.4	73.6	76.5	73.3
Conc.	60	60	60	60	60	60	60	60	60	60	60	60
Durab.	58.7	70.5	52.5	59.9	65.9	63.3	64.9	53.6	52.5	63.8	59.7	58.9

a. Have your computer program calculate a regression equation with durability predicted by concentration. Interpret the slope coefficient.
b. Find the coefficient of determination. What does it indicate about the predictive value of concentration?

11.91 In the regression model of Exercise 11.90, is the slope coefficient significantly different from 0 at $\alpha = .01$?

11.92 Obtain a plot of the data of Exercise 11.90, with durability on the vertical axis and concentration on the horizontal axis.
a. What does this plot indicate about the wisdom of using straight-line prediction?
b. What does this plot indicate about the correlation found in Exercise 11.90?

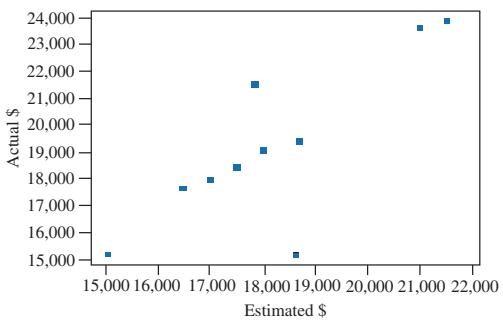
Bus. **11.93** A group of builders are considering a method for estimating the cost of constructing custom houses.

The builders used the method to estimate the cost of 10 "spec" houses that were built without a commitment from a customer. The builders obtained the actual costs (exclusive of land costs) of completing each house, to compare with the estimated costs.

"We went back to our accountant, who did a regression analysis of the data and gave us these results. The accountant says that the estimates are quite accurate, with an 80% correlation and a very low *p*-value. We're still pretty skeptical of whether this new method gives us decent estimates. We only clear a profit of about 10 percent, so a few bad estimates would hurt us. Can you explain to us what this output says about the estimating method?"

Write a brief, not-too-technical explanation for them. Focus on the builders' question about the accuracy of the estimates. A plot is shown here.

```
MTB > Regress 'Actual' on 1 variable 'Estimate'.
 The regression equation is
Actual = -34739 + 1.25 Estimate

Predictor      Coef       Stdev    t-ratio        p
Constant     -34739       60147      -0.58    0.579
Estimate     1.2474      0.3293       3.79    0.005

s = 19313      R-sq = 64.2%     R-sq(adj) = 59.7%

Analysis of Variance

SOURCE       DF         SS          MS         F       p
Regression    1 5350811136  5350811136     14.35   0.005
Error         8 2983948032   372993504
Total         9 8334758912

Unusual Observations
Obs. Estimate  Actual    Fit  Stdev.Fit   Residual   St.Resid   2   186200  152134
197531       6286    -45397      -2.49R

R denotes an obs. with a large st. resid.

MTB > Correlation 'Estimate' 'Actual'.

Correlation of Estimate and Actual = 0.801
```



CHAPTER 12

Multiple Regression and the General Linear Model

12.1 Introduction and Abstract of Research Study

In Chapter 11, we discussed the simplest type of regression model (simple linear regression) relating the response variable (also called the dependent variable) to a quantitative explanatory variable (also called the independent variable):

$$y = \beta_0 + \beta_1 x + \varepsilon$$

expected value of ε

In this chapter, we will generalize the above model to allow several explanatory variables and furthermore allow the explanatory variables to have categorical levels. In the simple linear model, the average value of ε (also called the **expected value of ε**) is restricted to be 0 for a given value of x. This restriction indicates that the average (expected) value of the response variable y for a given value of x is described by a straight line:

$$E(y) = \beta_0 + \beta_1 x$$

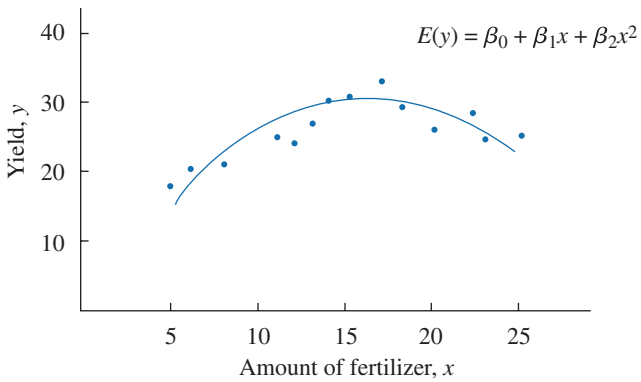
This model is very restrictive because in many research settings a straight line does not adequately represent the relationship between the response and explanatory variables.

For example, consider the data of Table 12.1, which gives the yields (in bushels) for 14 equal-sized plots planted in tomatoes for different levels of fertilization. It is evident from the scatterplot in Figure 12.1 that a linear equation will not adequately represent the relationship between yield and amount of fertilizer applied

TABLE 12.1
Yield of 14 equal-sized plots of tomato plantings for different amounts of fertilizer

Plot	Yield, y (in bushels)	Amount of Fertilizer, x (in pounds per plot)
1	24	12
2	18	5
3	31	15
4	33	17
5	26	20
6	30	14
7	20	6
8	25	23
9	25	11
10	27	13
11	21	8
12	29	18
13	29	22
14	26	25

FIGURE 12.1
Scatterplot of the yield versus fertilizer data in Table 12.1



to the plot. The reason for this is that, whereas a modest amount of fertilizer may well enhance the crop yield, too much fertilizer can be destructive.

A model for this physical situation might be

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Again with the assumption that $E(\varepsilon) = 0$, the expected value of y for a given value of x is

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

One such line is plotted in Figure 12.1, superimposed on the data of Table 12.1.

A general polynomial regression model relating a dependent variable y to a single quantitative independent variable x is given by

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

with

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$$

The choice of p and hence the choice of an appropriate regression model will depend on the experimental situation.

multiple regression model

The **multiple regression model,** which relates a response variable y to a set of k quantitative explanatory variables, is a direct extension of the polynomial regression model in one independent variable. The multiple regression model is expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

cross-product term

Any of the k explanatory variables may be powers of the independent variables, such as $x_3 = x_1^2$; a **cross-product term,** $x_4 = x_1 x_2$; a nonlinear function, such as $x_5 = \log(x_1)$; and so on. For the above definitions, we would have the following model:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 \log(x_1) + \varepsilon \end{aligned}$$

The only restriction is that no x_i is a perfect linear function of any other x_j. For example, $x_2 = 2 + 3x_1$ is not allowed.

first-order model

The simplest type of multiple regression equation is a **first-order model,** in which each of the independent variables appears, but there are no cross-product terms or terms in powers of the independent variables. For example, when three quantitative independent variables are involved, the first-order multiple regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

For these first-order models, we can attach some meaning to the βs. The parameter β_0 is the y-intercept, which represents the expected value of y when each x is zero. For cases in which it does not make sense to have each x be zero, β_0 (or its estimate) should be used only as part of the prediction equation and not given an interpretation by itself.

partial slopes

The other parameters ($\beta_1, \beta_2, \ldots, \beta_k$) in the multiple regression equation are sometimes called **partial slopes.** In linear regression, the parameter β_1 is the slope of the regression line, and it represents the expected change in y for a unit increase in x. In a first-order multiple regression model, β_1 represents the expected change in y for a unit increase in x_1 *when all other* x*s are held constant.* In general then, $\beta_j (j \neq 0)$ represents the expected change in y for a unit increase in x_j while holding all other xs constant. The usual assumptions for a multiple regression model are shown here.

DEFINITION 12.1

The **assumptions for multiple regression** are as follows:

1. The mathematical form of the relation is correct, so $E(\varepsilon_i) = 0$ for all i.
2. Var $(\varepsilon_i) = \sigma_\varepsilon^2$ for all i.
3. The ε_is are independent.
4. ε_i is normally distributed.

There is an additional assumption that is implied when we use a first-order multiple regression model. Because the expected change in y for a unit change

additive effects

in x_j is constant and does not depend on the value of any other x, we are in fact assuming that the effects of the independent variables are **additive.**

EXAMPLE 12.1

A brand manager for a new food product collected data on y = brand recognition (percent of potential consumers who can describe what the product is), x_1 = length in seconds of an introductory TV commercial, and x_2 = number of repetitions of the commercial over a 2-week period. What does the brand manager assume if a first-order model

$$\hat{y} = 0.31 + 0.042x_1 + 1.41x_2$$

is used to predict y?

Solution First, the manager assumes a straight-line, consistent rate of change. The manager assumes that a 1-second increase in length of the commercial will lead to a 0.042 percentage point increase in recognition, whether the increase is from, say, 10 to 11 seconds or from 59 to 60 seconds. Also, every additional repetition of the commercial is assumed to give a 1.41 percentage point increase in recognition, whether it is the second repetition or the twenty-second.

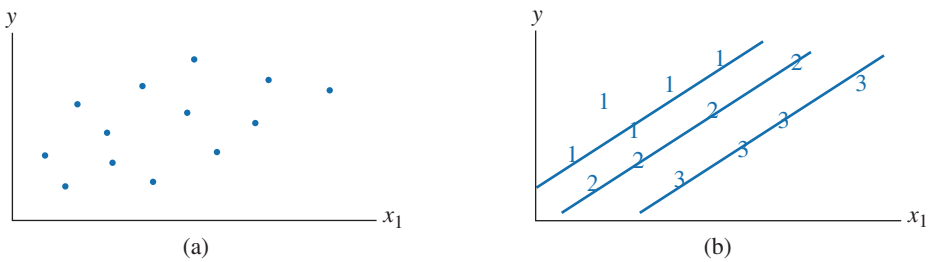
Second, there is a no-interaction assumption. The first-order model assumes that the effect of an additional repetition (that is, an increase in x_2) of a commercial of a given length (that is, holding x_1 constant) doesn't depend on *where* that length is held constant (at 10 seconds, 27 seconds, 60 seconds, whatever). ■

When might the additional assumption of additivity be warranted? Figure 12.2(a) shows a scatterplot of y versus x_1; Figure 12.2(b) shows the same plot with an ID attached to the different levels of a second independent variable x_2 (x_2 takes on the value of 1, 2, or 3). From Figure 12.2(a), we see that y is approximately linear in x_1. The parallel lines of Figure 12.2(b) corresponding to the three levels of the independent variable x_2 indicate that the expected change in y for a unit change in x_1 remains the same no matter which level of x_2 is used. These data suggest that the effects of x_1 and x_2 are additive; hence, a first-order model of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ is appropriate.

interaction

Figure 12.3 displays a situation in which **interaction** is present between the variables x_1 and x_2. The nonparallel lines in Figure 12.3 indicate that the change in the expected value of y for a unit change in x_1 varies depending on the value of x_2. In particular, it can be noted that when $x_1 = 10$, there is almost no difference in the expected value of y for the three values of x_2. However, when $x_1 = 50$, the

FIGURE 12.2
(a) Scatterplot of y versus x_1. (b) Scatterplot of y versus x_1, indicating additivity of effects for x_1 and x_2.



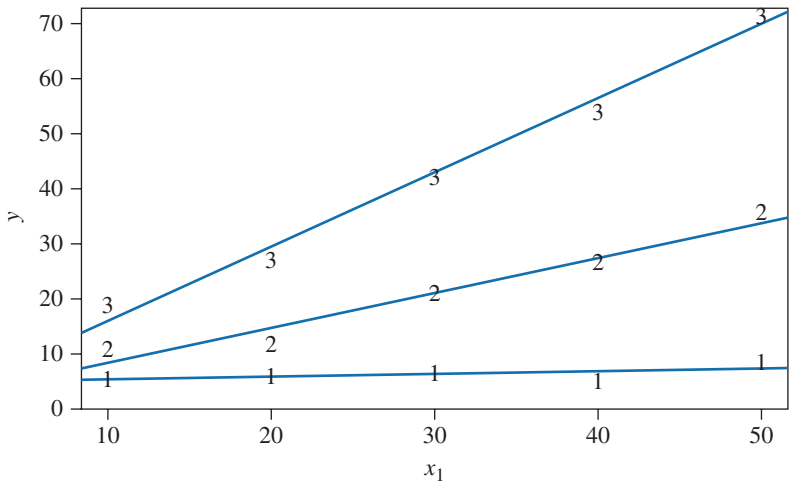


FIGURE 12.3
Scatterplot of y versus x_1 at three levels of x_2

expected value of y when $x_2 = 3$ is much larger than the values of the expected value of y for $x_2 = 2$ and $x_2 = 1$. Thus, the rate of change in the expected value of y has increased much more rapidly for $x_2 = 3$ than it does for $x_2 = 1$. When this type of relationship exists, the explanatory variables are said to interact. A first-order model, which assumes no interaction, would not be appropriate in the situation depicted in Figure 12.3. At the very least, it is necessary to include a cross-product term (x_1x_2) in the model.

The simplest model allowing for interaction between x_1 and x_2 is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Note that for a given value of x_2 (say, $x_2 = 2$), the expected value of y is

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x_1 + \beta_2(2) + \beta_3 x_1(2) \\ &= (\beta_0 + 2\beta_2) + (\beta_1 + 2\beta_3)x_1 \end{aligned}$$

Here the intercept and slope are $(\beta_0 + 2\beta_2)$ and $(\beta_1 + 2\beta_3)$, respectively. The corresponding intercept and slope for $x_2 = 3$ can be shown to be $(\beta_0 + 3\beta_2)$ and $(\beta_1 + 3\beta_3)$. Clearly, the slopes of the two regression lines are not the same, and, hence, we have nonparallel lines.

Not all experiments can be modeled using a first-order multiple regression model. For these situations, in which a higher-order multiple regression model may be appropriate, it will be more difficult to assign a literal interpretation to the βs because of the presence of terms that contain cross-products or powers of the independent variables. Our focus will be on finding a multiple regression model that provides a good fit to the sample data, not on interpreting individual βs, except as they relate to the overall model.

The models that we have described briefly have been for regression problems for which the experimenter is interested in developing a model to relate a response to one or more *quantitative* independent variables. The problem of modeling an experimental situation is not restricted to the quantitative independent-variable case.

Consider the problem of writing a model for an experimental situation in which a response y is related to a set of *qualitative* independent variables or to both quantitative and qualitative independent variables. For the first situation (relating

y to one or more qualitative independent variables), let us suppose that we want to compare the average number of lightning discharges per minute for a storm, as measured from two different tracking posts located 30 miles apart. If we let y denote the number of discharges recorded on an oscilloscope during a 1-minute period, we could write the following two models:

$$\text{For tracking post 1: } y = \mu_1 + \varepsilon$$
$$\text{For tracking post 2: } y = \mu_2 + \varepsilon$$

Thus, we assume that observations at tracking post 1 randomly "fluctuate" about a population mean μ_1. Similarly, at tracking post 2, observations differ from a population mean μ_2 by a random amount ε. These two models are not new and could have been used to describe observations when comparing two population means in Chapter 6. What is new is that we can combine these two models into a single model of the form

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

where β_0 and β_1 are unknown parameters, ε is a random error term, and x_1 is a **dummy variable** with the following interpretation. We let

dummy variable

$$x_1 = 1 \text{ if an observation is obtained from tracking post 2}$$
$$x_1 = 0 \text{ if an observation is obtained from tracking post 1}$$

For observations obtained from tracking post 1, we substitute $x_1 = 0$ into our model to obtain

$$y = \beta_0 + \beta_1(0) + \varepsilon = \beta_0 + \varepsilon$$

Hence, $\beta_0 = \mu_1$, the population mean for observations from tracking post 1. Similarly, by substituting $x_1 = 1$ in our model, the equation for observations from tracking post 2 is

$$y = \beta_0 + \beta_1(1) + \varepsilon = \beta_0 + \beta_1 + \varepsilon$$

Because $\beta_0 = \mu_1$ and $\beta_0 + \beta_1$ must equal μ_2, we have $\beta_1 = \mu_2 - \mu_1$, the difference in means between observations from tracking posts 2 and 1.

This model, $y = \beta_0 + \beta_1 x_1 + \varepsilon$, which relates y to the qualitative independent variable tracking post, can be extended to a situation in which the qualitative variable has more than two levels. We do this by using more than one dummy variable. Consider an experiment in which we're interested in four levels of qualitative variables. We call these levels **treatments.** We could write the model

treatments

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

$$x_1 = 1 \text{ if treatment 2}, \quad x_1 = 0 \text{ otherwise}$$
$$x_2 = 1 \text{ if treatment 3}, \quad x_2 = 0 \text{ otherwise}$$
$$x_3 = 1 \text{ if treatment 4}, \quad x_3 = 0 \text{ otherwise}$$

To interpret the βs in this equation, it is convenient to construct a table of the expected values. Because ε has expectation zero, the general expression for the expected value of y is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

TABLE 12.2
Expected values for an experiment with four treatments

Treatment			
1	**2**	**3**	**4**
$E(y) = \beta_0$	$E(y) = \beta_0 + \beta_1$	$E(y) = \beta_0 + \beta_2$	$E(y) = \beta_0 + \beta_3$

The expected value for observations on treatment 1 is found by substituting $x_1 = 0, x_2 = 0$, and $x_3 = 0$; after this substitution, we find $E(y) = \beta_0$. The expected value for observations on treatment 2 is found by substituting $x_1 = 1, x_2 = 0$, and $x_3 = 0$ into the $E(y)$ formula; this substitution yields $E(y) = \beta_0 + \beta_1$. Substitutions of $x_1 = 0, x_2 = 1, x_3 = 0$ and $x_1 = 0, x_2 = 0, x_3 = 1$ yield expected values for treatments 3 and 4, respectively. These expected values are summarized in Table 12.2.

If we identify the mean of treatment 1 as μ_1, the mean of treatment 2 as μ_2, and so on, then from Table 12.2 we have

$$\mu_1 = \beta_0 \quad \mu_2 = \beta_0 + \beta_1 \quad \mu_3 = \beta_0 + \beta_2 \quad \mu_4 = \beta_0 + \beta_3$$

Solving these equations for the βs, we have

$$\beta_0 = \mu_1 \quad \beta_1 = \mu_2 - \mu_1 \quad \beta_2 = \mu_3 - \mu_1 \quad \beta_3 = \mu_4 - \mu_1$$

Any comparison among the treatment means can be phrased in terms of the βs. For example, the comparison $\mu_4 - \mu_3$ could be written as $\beta_3 - \beta_2$, and $\mu_3 - \mu_2$ could be written as $\beta_2 - \beta_1$.

EXAMPLE 12.2

An industrial engineer is designing a simulation model to generate the time needed to retrieve parts from a warehouse under four different automated retrieval systems. Suppose the mean times as provided by the companies producing the systems are $\mu_1 = 7, \mu_2 = 9, \mu_3 = 6$, and $\mu_4 = 15$. The engineer uses the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

$x_1 = 1$ if system 2 is used, $x_1 = 0$ otherwise
$x_2 = 1$ if system 3 is used, $x_2 = 0$ otherwise
$x_3 = 1$ if system 4 is used, $x_3 = 0$ otherwise

Using the values of the retrieval means, determine the values for $\beta_0, \beta_1, \beta_2$, and β_3 to be used in the above model.

Solution Based on what we saw in Table 12.2, we know that

$$\beta_0 = \mu_1 \quad \beta_1 = \mu_2 - \mu_1 \quad \beta_2 = \mu_3 - \mu_1 \quad \beta_3 = \mu_4 - \mu_1$$

Using the known values for μ_1, μ_2, μ_3, and μ_4, it follows that

$$\beta_0 = 7 \quad \beta_1 = 9 - 7 = 2 \quad \beta_2 = 6 - 7 = -1 \quad \beta_3 = 15 - 7 = 8 \blacksquare$$

EXAMPLE 12.3

Refer to Example 12.2. Express $\mu_3 - \mu_2$ and $\mu_3 - \mu_4$ in terms of the βs. Check your findings by substituting values for the βs.

Solution Using the relationship between the βs and the μs, we can see that

$$\beta_2 - \beta_1 = (\mu_3 - \mu_1) - (\mu_2 - \mu_1) = \mu_3 - \mu_2$$

and

$$\beta_2 - \beta_3 = (\mu_3 - \mu_1) - (\mu_4 - \mu_1) = \mu_3 - \mu_4$$

Substituting computed values for the βs, we have

$$\beta_2 - \beta_1 = -1 - (2) = -3$$

and

$$\beta_2 - \beta_3 = -1 - (8) = -9$$

These computed values are identical to the "known" differences for $\mu_3 - \mu_2$ and $\mu_3 - \mu_4$, respectively. ■

EXAMPLE 12.4

Use dummy variables to write the model for an experiment with t treatments. Identify the βs.

Solution We can write the model in the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{t-1} x_{t-1} + \varepsilon$$

where

$$\begin{aligned} x_1 &= 1 \text{ if treatment 2}, & x_1 &= 0 \text{ otherwise} \\ x_2 &= 1 \text{ if treatment 3}, & x_2 &= 0 \text{ otherwise} \\ &\vdots & &\vdots \\ x_{t-1} &= 1 \text{ if treatment } t, & x_{t-1} &= 0 \text{ otherwise} \end{aligned}$$

The table of expected values would be as shown in Table 12.3, from which we obtain

$$\begin{aligned} \beta_0 &= \mu_1 \\ \beta_1 &= \mu_2 - \mu_1 \\ &\vdots \\ \beta_{t-1} &= \mu_t - \mu_1 \end{aligned}$$

TABLE 12.3 Expected values

Treatment			
1	**2**	**...**	***t***
$E(y) = \beta_0$	$E(y) = \beta_0 + \beta_1$	...	$E(y) = \beta_0 + \beta_{t-1}$

In the procedure just described, we have a response related to the qualitative variable "treatments," and for t levels of the treatments, we enter $(t - 1)$ βs into our model, using dummy variables. ■

More will be said about the use of the models for more than one qualitative independent variable in Chapters 14 and 15, where we consider the analysis of variance for several different experimental designs. In Chapter 16, we will also consider models in which there are both quantitative and qualitative variables.

Abstract of Research Study: Evaluation of the Performance of an Electric Drill

In recent years, there have been numerous reports of homeowners encountering problems with electric drills. The drills would tend to overheat when under strenuous usage. A consumer product testing laboratory has selected a variety of brands of electric drills to determine what types of drills are most and least likely to overheat under specified conditions. After a careful evaluation of the differences in the designs of the drills, the engineers selected three design factors for use in comparing the resistance of the drills to overheating. The design factors were the thickness of the insulation around the motor, the quality of the wire used in the drill's motor, and the size of the vents in the body of the drill.

The engineers designed a study taking into account various combinations of the three design factors. There were five levels of the thickness of the insulation, three levels of the quality of the wire used in the motor, and three sizes for the vents in the drill body. Thus, the engineers had potentially 45 ($5 \times 3 \times 3$) uniquely designed drills. However, each of these 45 drills would have differences with respect to other factors that may vary their performance. Thus, the engineers selected 10 drills of each of the 45 designs. Another factor that may vary the results of the study is the conditions under which each of the drills is tested. The engineers selected two "torture tests" that they felt reasonably represented the types of conditions under which overheating occurred. The 10 drills were then randomly assigned to one of the two torture tests. At the end of the test, the temperature of the drill was recorded. The mean temperature of the 5 drills was the response variable of interest to the engineers. A second response variable was the logarithm of the sample variance of the 5 drills. This response variable measures the degree to which the 5 drills produced a consistent temperature under each of the torture tests. The goal of the study was to determine which combination of the design factors of the drills produced the smallest values of both response variables. Thus, they would obtain a design for a drill having minimum mean temperature and a design that produced drills for which an individual drill was most likely to produce a temperature closest to the mean temperature. An analysis of the 90 drill responses in order to determine the "best" design for the drill is given in Section 12.10. The data from this study are given in Table 12.4 with the following notation:

AVTEM: mean temperature for the five drills under a given torture test
LOGV: logarithm of the variance of the temperatures of the five drills
IT: the thickness of the insulation within the drill (IT = 2, 3, 4, 5, or 6)
QW: an assessment of quality of the wire used in the drill motor (QW = 6, 7, or 8)
VS: the size of the vent used in the motor (VS = 10, 11, or 12)
$I2 = (IT - \text{mean IT})^2$, $Q2 = (QW - \text{mean QW})^2$, $V2 = (VS - \text{mean VS})^2$
TEST: the type of torture test used

TABLE 12.4
Drill performance data

AVTEM	LOGV	IT	QW	VS	I2	Q2	V2	Test	AVTEM	LOGV	IT	QW	VS	I2	Q2	V2	Test
185	3.6	2	6	10	4	1	1	1	168	3.4	4	7	11	0	0	0	2
176	3.7	2	6	10	4	1	1	2	160	2.9	4	7	12	0	0	1	1
177	3.6	2	6	11	4	1	0	1	154	3.1	4	7	12	0	0	1	2
184	3.7	2	6	11	4	1	0	2	169	2.8	4	8	10	0	1	1	1
178	3.6	2	6	12	4	1	1	1	156	2.9	4	8	10	0	1	1	2
169	3.4	2	6	12	4	1	1	2	168	2.7	4	8	11	0	1	0	1
185	3.2	2	7	10	4	0	1	1	161	2.7	4	8	11	0	1	0	2
184	3.2	2	7	10	4	0	1	2	156	2.6	4	8	12	0	1	1	1
180	3.2	2	7	11	4	0	0	1	158	2.7	4	8	12	0	1	1	2
184	3.5	2	7	11	4	0	0	2	164	3.7	5	6	10	1	1	1	1
179	3.0	2	7	12	4	0	1	1	163	3.7	5	6	10	1	1	1	2
173	3.2	2	7	12	4	0	1	2	161	3.7	5	6	11	1	1	0	1
179	2.9	2	8	10	4	1	1	1	158	3.4	5	6	11	1	1	0	2
185	2.7	2	8	10	4	1	1	2	154	3.4	5	6	12	1	1	1	1
180	2.8	2	8	11	4	1	0	1	162	3.7	5	6	12	1	1	1	2
180	2.7	2	8	11	4	1	0	2	163	2.8	5	7	10	1	0	1	1
169	2.9	2	8	12	4	1	1	1	166	3.0	5	7	10	1	0	1	2
177	2.8	2	8	12	4	1	1	2	159	3.3	5	7	11	1	0	0	1
172	3.6	3	6	10	1	1	1	1	156	3.3	5	7	11	1	0	0	2
171	3.9	3	6	10	1	1	1	2	152	3.3	5	7	12	1	0	1	1
172	3.8	3	6	11	1	1	0	1	150	3.3	5	7	12	1	0	1	2
167	3.6	3	6	11	1	1	0	2	165	2.9	5	8	10	1	1	1	1
165	3.3	3	6	12	1	1	1	1	156	2.7	5	8	10	1	1	1	2
159	3.4	3	6	12	1	1	1	2	155	2.8	5	8	11	1	1	0	1
169	3.0	3	7	10	1	0	1	1	155	3.2	5	8	11	1	1	0	2
174	3.3	3	7	10	1	0	1	2	149	2.6	5	8	12	1	1	1	1
163	3.3	3	7	11	1	0	0	1	152	2.9	5	8	12	1	1	1	2
170	3.3	3	7	11	1	0	0	2	165	3.4	6	6	10	4	1	1	1
169	3.2	3	7	12	1	0	1	1	160	3.7	6	6	10	4	1	1	2
163	3.2	3	7	12	1	0	1	2	157	3.7	6	6	11	4	1	0	1
178	2.7	3	8	10	1	1	1	1	149	3.7	6	6	11	4	1	0	2
165	2.7	3	8	10	1	1	1	2	149	3.8	6	6	12	4	1	1	1
167	2.8	3	8	11	1	1	0	1	145	3.7	6	6	12	4	1	1	2
171	2.8	3	8	11	1	1	0	2	154	3.4	6	7	10	4	0	1	1
166	2.9	3	8	12	1	1	1	1	153	3.2	6	7	10	4	0	1	2
166	2.7	3	8	12	1	1	1	2	150	3.0	6	7	11	4	0	0	1
161	3.7	4	6	10	0	1	1	1	156	3.1	6	7	11	4	0	0	2
162	3.7	4	6	10	0	1	1	2	146	3.2	6	7	12	4	0	1	1
169	3.4	4	6	11	0	1	0	1	153	3.3	6	7	12	4	0	1	2
162	3.7	4	6	11	0	1	0	2	161	2.8	6	8	10	4	1	1	1
159	3.5	4	6	12	0	1	1	1	160	2.9	6	8	10	4	1	1	2
168	3.4	4	6	12	0	1	1	2	156	2.9	6	8	11	4	1	0	1
169	3.1	4	7	10	0	0	1	1	150	2.7	6	8	11	4	1	0	2
165	3.2	4	7	10	0	0	1	2	149	2.9	6	8	12	4	1	1	1
163	3.2	4	7	11	0	0	0	1	151	2.8	6	8	12	4	1	1	2

12.2 The General Linear Model

general linear model

It is important at this point to recognize that a single general model can be used for multiple regression models in which a response is related to a set of quantitative independent variables and for models that relate y to a set of qualitative independent variables. This model, called the **general linear model,** has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

For multiple regression models, the xs represent quantitative independent variables (such as weight or amount of water), independent variables raised to powers, and cross-product terms involving the independent variables. We discussed a few regression models in Section 12.1; more about the use of the general linear model in regression will be discussed in the remainder of this chapter and in Chapter 13.

When y is related to a set of qualitative independent variables, the xs of the general linear model represent dummy variables (coded 0 and 1) or products of dummy variables. We discussed how to use dummy variables for representing y in terms of a single qualitative variable in Section 12.1; the same approach can be used to relate y to more than one qualitative independent variable. This will be discussed in Chapter 14, where we present more analysis of variance techniques.

The general linear model can also be used for the case in which y is related to both qualitative and quantitative independent variables. A particular example of this is discussed in Section 12.7, and other applications are presented in Chapter 16.

Why is this model called the general *linear* model, especially as it can be used for polynomial models? The word *linear* in the general linear model refers to how the βs are entered in the model, not to how the independent variables appear in the model. A general linear model is linear (used in the usual algebraic sense) in the βs.

That is, the βs do not appear as an exponent or as the argument of a nonlinear function. Examples of models which are not linear models include

- $y = \beta_1 x_1 e^{\beta_2 x_2} + \varepsilon$

  (nonlinear because β_2 appears as an exponent).
- $y = \beta_1 \text{cosine}(\beta_2 x_2) + \varepsilon$

  (nonlinear because β_2 appears as an argument of the cosine function).

The following two models will be referred to as linear models, even though they are not linear in the explanatory variable, because they are linear in βs:

- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

  β_0, β_1, and β_2 appear as coefficients in a quadratic model in x.
- $y = \beta_0 + \beta_1 \text{sine}(x_1) + \beta_2 \log(x_2) + \varepsilon$

  β_0, β_1, and β_2 appear as coefficients in a model involving functions of the two explanatory variables x_1 and x_2.

Why are we discussing the general linear model now? The techniques that we will develop in this chapter for making inferences about a single β, a set of βs, and $E(y)$ in multiple regression are those that apply to any general linear model. Thus, using general linear model techniques, we have a common thread to inferences about multiple regression (Chapters 12 and 13) and the analysis

of variance (Chapters 14 through 18). As you study these seven chapters, try whenever possible to make the connection back to a general linear model; we'll help you with this connection. For Sections 12.3 through 12.10 of this chapter, we will concentrate on multiple regression, which is a special case of a general linear model.

12.3 Estimating Multiple Regression Coefficients

The multiple regression model relates a response y to a set of quantitative independent variables. For a random sample of n measurements, we can write the ith observation as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad (i = 1, 2, \ldots, n; n > k)$$

where $x_{i1}, x_{i2}, \ldots, x_{ik}$ are the settings of the quantitative independent variables corresponding to the observation y_i.

To find least-squares estimates for $\beta_0, \beta_1, \ldots,$ and β_k in a multiple regression model, we follow the same procedure that we did for a linear regression model in Chapter 11. We obtain a random sample of n observations; we find the least-squares prediction equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

by choosing $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ to minimize SS(Residual) $= \sum_i (y_i - \hat{y}_i)^2$. However, although it was easy to write down the solutions to $\hat{\beta}_0$ and $\hat{\beta}_1$ for the linear regression model,

$$y = \beta_0 + \beta_1 x + \varepsilon$$

we must find the estimates for $\beta_0, \beta_1, \ldots, \beta_k$ by solving a set of simultaneous equations, called the *normal equations,* shown in Table 12.5.

TABLE 12.5
Normal equations for a multiple regression model

	y_i	$\hat{\beta}_0$	$x_{i1}\hat{\beta}_1$	...	$x_{ik}\hat{\beta}_k$
1	$\sum y_i =$	$n\hat{\beta}_0$ +	$\sum x_{i1}\hat{\beta}_1$	+ ⋯ +	$\sum x_{ik}\hat{\beta}_k$
x_{i1}	$\sum x_{i1}y_i =$	$\sum x_{i1}\hat{\beta}_0$ +	$\sum x_{i1}^2\hat{\beta}_1$	+ ⋯ +	$\sum x_{i1}x_{ik}\hat{\beta}_k$
⋮	⋮				
x_{ik}	$\sum x_{ik}y_i =$	$\sum x_{ik}\hat{\beta}_0$ +	$\sum x_{ik}x_{i1}\hat{\beta}_1$	+ ⋯ +	$\sum x_{ik}^2\hat{\beta}_k$

Note the pattern associated with these equations. By labeling the rows and columns as we have done, we can obtain any term in the normal equations by multiplying the row and column elements and summing. For example, the last term in the second equation is found by multiplying the row element (x_{i1}) by the column element $(x_{ik}\hat{\beta}_k)$ and summing; the resulting term is $\sum x_{i1}x_{ik}\hat{\beta}_k$. Because all terms in the normal equations can be formed in this way, it is fairly simple to write down the equations to be solved to obtain the least-squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. The solution to these equations is not necessarily trivial; that's why we'll enlist the help of various statistical software packages for their solution.

EXAMPLE 12.5

An experiment was conducted to investigate the weight loss of a compound for different amounts of time the compound was exposed to the air. Additional information was also available on the humidity of the environment during exposure. The complete data are presented in Table 12.6.

TABLE 12.6
Weight loss, exposure time, and relative humidity data

Weight Loss, y (pounds)	Exposure Time, x_1 (hours)	Relative Humidity, x_2
4.3	4	.20
5.5	5	.20
6.8	6	.20
8.0	7	.20
4.0	4	.30
5.2	5	.30
6.6	6	.30
7.5	7	.30
2.0	4	.40
4.0	5	.40
5.7	6	.40
6.5	7	.40

a. Set up the normal equations for this regression problem if the assumed model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where x_1 is exposure time and x_2 is relative humidity.

b. Use the computer output shown here to determine the least-squares estimates of β_0, β_1, and β_2. Predict weight loss for 6.5 hours of exposure and a relative humidity of .35.

OUTPUT FOR EXAMPLE 12.5

OBS	WT_LOSS	TIME	HUMID
1	4.3	4.0	0.20
2	5.5	5.0	0.20
3	6.8	6.0	0.20
4	8.0	7.0	0.20
5	4.0	4.0	0.30
6	5.2	5.0	0.30
7	6.6	6.0	0.30
8	7.5	7.0	0.30
9	2.0	4.0	0.40
10	4.0	5.0	0.40
11	5.7	6.0	0.40
12	6.5	7.0	0.40
13	.	6.5	0.35

Dependent Variable: WT_LOSS WEIGHT LOSS

```
Analysis of Variance

                          Sum of          Mean
Source          DF       Squares        Square     F Value      Prob>F

Model            2      31.12417      15.56208     104.133      0.0001
Error            9       1.34500       0.14944
C Total         11      32.46917

    Root MSE       0.38658    R-square      0.9586
    Dep Mean       5.50833    Adj R-sq      0.9494
    C.V.           7.01810

Parameter Estimates

                  Parameter      Standard   T for H0:
Variable  DF       Estimate         Error  Parameter=0   Prob > |T|

INTERCEP   1       0.666667    0.69423219        0.960        0.3620
TIME       1       1.316667    0.09981464       13.191        0.0001
HUMID      1      -8.000000    1.36676829       -5.853        0.0002

OBS    WT_LOSS      PRED      RESID    L95MEAN    U95MEAN

  1        4.3     4.33333   -0.03333    3.80985    4.85682
  2        5.5     5.65000   -0.15000    5.23519    6.06481
  3        6.8     6.96667   -0.16667    6.55185    7.38148
  4        8.0     8.28333   -0.28333    7.75985    8.80682
  5        4.0     3.53333    0.46667    3.11091    3.95576
  6        5.2     4.85000    0.35000    4.57346    5.12654
  7        6.6     6.16667    0.43333    5.89012    6.44321
  8        7.5     7.48333    0.01667    7.06091    7.90576
  9        2.0     2.73333   -0.73333    2.20985    3.25682
 10        4.0     4.05000   -0.05000    3.63519    4.46481
 11        5.7     5.36667    0.33333    4.95185    5.78148
 12        6.5     6.68333   -0.18333    6.15985    7.20682
 13          .     6.42500          .    6.05269    6.79731

Sum of Residuals                   0
Sum of Squared Residuals      1.3450
Predicted Resid SS (Press)    2.6123
```

Solution

a. The three normal equations for this model are shown in Table 12.7.

TABLE 12.7 Normal equations for Example 12.5

	y_i	$\hat{\beta}_0$	$x_{i1}\hat{\beta}_1$	$x_{i2}\hat{\beta}_2$
1	$\sum y_i$ =	$n\hat{\beta}_0$ +	$\sum x_{i1}\hat{\beta}_1$ +	$\sum x_{i2}\hat{\beta}_2$
x_{i1}	$\sum x_{i1}y_i$ =	$\sum x_{i1}\hat{\beta}_0$ +	$\sum x_{i1}^2\hat{\beta}_1$ +	$\sum x_{i1}x_{i2}\hat{\beta}_2$
x_{i2}	$\sum x_{i2}y_i$ =	$\sum x_{i2}\hat{\beta}_0$ +	$\sum x_{i2}x_{i1}\hat{\beta}_1$ +	$\sum x_{i2}^2\hat{\beta}_2$

For these data, we have

$$\sum y_i = 66.10 \qquad \sum x_{i1} = 66 \qquad \sum x_{i2} = 3.60$$

$$\sum x_{i1}y_i = 383.3 \qquad \sum x_{i2}y_i = 19.19 \qquad \sum x_{i1}x_{i2} = 19.8$$

$$\sum x_{i1}^2 = 378 \qquad \sum x_{i2}^2 = 1.16$$

Substituting these values into the normal equation yields the result shown here:

$$66.1 = 12\hat{\beta}_0 + 66\hat{\beta}_1 + 3.6\hat{\beta}_2$$
$$383.3 = 66\hat{\beta}_0 + 378\hat{\beta}_1 + 19.8\hat{\beta}_2$$
$$19.19 = 3.6\hat{\beta}_0 + 19.8\hat{\beta}_1 + 1.16\hat{\beta}_2$$

b. The normal equations of part (a) could be solved to determine $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\beta}_2$. The solution would agree with that shown here in the output. The least-squares prediction equation is

$$\hat{y} = 0.667 + 1.317x_1 - 8.000x_2$$

where x_1 is exposure time and x_2 is relative humidity. Substituting $x_1 = 6.5$ and $x_2 = .35$, we have

$$\hat{y} = 0.667 + 1.317(6.5) - 8.000(.35) = 6.428$$

This value agrees with the predicted value shown as observation 13 in the output, except for rounding errors. ■

There are many software programs that provide the calculations to obtain least-squares estimates for parameters in the general linear model (and hence for multiple regression). The output of such programs typically has a list of variable names, together with the estimated partial slopes, labeled COEFFICIENTS (or ESTIMATES or PARAMETERS). The intercept term $\hat{\beta}_0$ is usually called INTERCEPT (or CONSTANT); sometimes it is shown along with the slopes but with no variable name.

EXAMPLE 12.6

A kinesiologist is investigating measures of the physical fitness of persons entering 10-kilometer races. A major component of overall fitness is cardiorespiratory capacity as measured by maximal oxygen uptake. Direct measurement of maximal oxygen is expensive and thus is difficult to apply to large groups of individuals in a timely fashion. The researcher wanted to determine if a prediction of maximal oxygen uptake can be obtained from a prediction equation using easily measured explanatory variables from the runners. In a preliminary study, the kinesiologist randomly selects 54 males and obtains the following data for the variables

y = maximal oxygen uptake (in liters per minute)

x_1 = weight (in kilograms)

x_2 = age (in years)

x_3 = time necessary to walk 1 mile (in minutes)

x_4 = heart rate at end of the walk (in beats per minute)

The data shown in Table 12.8 were simulated from a model that is consistent with information given in the article ***"Validation of the Rockport Fitness Walking Test in College Males and Females"*** **[Research Quarterly for Exercise and Sport *(1994) 65: 152–158].***

TABLE 12.8
Fitness walking test data

	Subject											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
y	1.5	2.1	1.8	2.2	2.2	2.0	2.1	1.9	2.8	1.9	2.0	2.7
x_1	139.8	143.3	154.2	176.6	154.3	185.4	177.9	158.8	159.8	123.9	164.2	146.3
x_2	19.1	21.1	21.2	23.2	22.4	22.1	21.6	19.0	20.9	22.0	19.5	19.8
x_3	18.1	15.3	15.3	17.7	17.1	16.4	17.3	16.8	15.5	13.8	17.0	13.8
x_4	133.6	144.6	164.6	139.4	127.3	137.3	144.0	141.4	127.7	124.2	135.7	116.1
	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
y	2.4	2.3	2.0	1.7	2.3	0.9	1.2	1.9	0.8	2.2	2.3	1.7
x_1	172.6	147.5	163.0	159.8	162.7	133.3	142.8	146.6	141.6	158.9	151.9	153.3
x_2	20.7	21.0	21.2	20.4	20.0	21.1	22.6	23.0	22.1	22.8	21.8	20.0
x_3	16.8	15.3	14.2	16.8	16.6	17.5	18.0	15.7	19.1	13.4	13.6	16.1
x_4	109.0	131.0	143.3	156.6	120.1	131.8	149.4	106.9	135.6	164.6	162.6	134.8
	25	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
y	1.6	1.6	2.8	2.7	1.3	2.1	2.5	1.5	2.4	2.3	1.9	1.5
x_1	144.6	133.3	153.6	158.6	108.4	157.4	141.7	151.1	149.5	144.3	166.6	153.6
x_2	22.9	22.9	19.4	21.0	21.1	20.1	19.8	21.8	20.5	21.0	21.4	20.8
x_3	15.8	18.2	13.3	14.9	16.7	15.7	13.5	18.8	14.9	17.2	17.4	16.4
x_4	154.0	120.7	151.9	133.6	142.8	168.2	120.5	135.6	119.5	119.0	150.8	144.0
	37	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**
y	2.4	2.3	1.7	2.0	1.9	2.3	2.1	2.2	1.8	2.1	2.2	1.3
x_1	144.1	148.7	159.9	162.8	145.7	156.7	162.3	164.7	134.4	160.1	143.0	141.6
x_2	20.3	19.1	19.6	21.3	20.0	19.2	22.1	19.1	20.9	21.1	20.5	21.7
x_3	13.3	15.4	17.4	16.2	18.6	16.4	19.0	17.1	15.6	14.2	17.1	14.5
x_4	124.7	154.4	136.7	152.4	133.6	113.2	81.6	134.8	130.4	162.1	144.7	163.1
	49	**50**	**51**	**52**	**53**	**54**						
y	2.5	2.2	1.4	2.2	2.5	1.8						
x_1	152.0	187.1	122.9	157.1	155.1	133.6						
x_2	20.8	21.5	22.6	23.4	20.8	22.5						
x_3	17.3	14.6	18.6	14.2	16.0	15.4						
x_4	137.1	156.0	127.2	121.4	155.3	140.4						

The data in Table 12.8 were analyzed using Minitab software. Identify the least-squares estimators of the intercept and partial slopes.

```
Regression Analysis:  y versus wgt, age, time, pulse

The regression equation is
y = 5.59 + 0.0129 wgt - 0.0830 age - 0.158 time - 0.00911 pulse


Predictor      Coef   SE Coef      T      P  VIF
Constant      5.588     1.030   5.43  0.000
wgt        0.012906  0.002827   4.57  0.000  1.0
age        -0.08300   0.03484  -2.38  0.021  1.0
time       -0.15817   0.02658  -5.95  0.000  1.1
pulse    -0.009114  0.002507  -3.64  0.001  1.1
```

Solution The least-squares estimator of the intercept, $\hat{\beta}_0$, is 5.588 and is labeled as *Constant*. The least-squares estimators of the four partial slopes—.012906, −.08300, −.15817, and −.009114—are associated with the explanatory variables, weight (wgt), age of subject (age), time to complete 1-mile walk (time), and heart rate at end of walk (pulse), respectively. The labels for the estimators of the intercept and partial slopes vary across the various software programs. ■

The coefficient of an independent variable x_j in a multiple regression equation does not, in general, equal the coefficient that would apply to that variable in a simple linear regression. In multiple regression, the coefficient refers to the effect of changing that x_j variable while other independent variables stay constant. In simple linear regression, all other potential independent variables are ignored. If other independent variables are correlated with x_j (and therefore don't tend to stay constant while x_j changes), simple linear regression with x_j as the only independent variable captures not only the direct effect of changing x_j but also the indirect effect of the associated changes in other xs. In multiple regression, by holding the other xs constant, we eliminate that indirect effect.

EXAMPLE 12.7

Refer to the data in Example 12.6. A multiple regression model was run using the SPSS software, yielding the output shown in Table 12.9.

TABLE 12.9
SPSS output for multiple regression model of Example 12.6

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	5.588	1.030		5.426	.000
	wgt	.013	.003	.426	4.565	.000
	age	−.083	.035	−.221	−2.382	.021
	time	−.158	.027	−.570	−5.950	.000
	pulse	−.009	.003	−.350	−3.636	.001

a. Dependent variable: *y*

Next, a simple linear regression (one-explanatory-variable) model was run using just the variable x_4, *pulse,* yielding the output in Table 12.10.

TABLE 12.10
SPSS output for a simple linear regression model relating x_4 to y

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.545	.494		5.153	.000
	pulse	−.004	.004	−.152	−1.111	.272

a. Dependent variable: *y*

Compare the coefficients of *pulse* in the two models. Explain why the two coefficients differ.

Solution In the multiple regression model, the least-squares regression model was estimated to be

$$y = 5.588 + .013x_1 - .083x_2 - .158x_3 - .009x_4$$

In the simple linear regression model, the least-squares regression model was estimated to be

$$y = 2.545 - .004x_4$$

The difference occurs because the four explanatory variables are correlated, as displayed in the output in Table 12.11.

TABLE 12.11
Correlations between the variables in Example 12.6

Correlations

		y	wgt	age	time	pulse
y	Pearson Correlation	1	.414**	−.288*	−.506**	−.152
	Sig. (2-tailed)		.002	.035	.000	.272
	N	54	54	54	54	54
wgt	Pearson Correlation	.414**	1	−.074	−.022	.116
	Sig. (2-tailed)	.002		.596	.873	.404
	N	54	54	54	54	54
age	Pearson Correlation	−.288*	−.074	1	.069	−.013
	Sig. (2-tailed)	.035	.596		.619	.926
	N	54	54	54	54	54
time	Pearson Correlation	−.506**	−.022	.069	1	−.255
	Sig. (2-tailed)	.000	.873	.619		.063
	N	54	54	54	54	54
pulse	Pearson Correlation	−.152	.116	−.013	−.255	1
	Sig. (2-tailed)	.272	.404	.926	.063	
	N	54	54	54	54	54

** Correlation is significant at the .01 level (2-tailed).

* Correlation is significant at the .05 level (2-tailed).

In the simple linear regression model, $\hat{\beta}_1 = -.004$ represents a decrease of .004 liters per minute in y, maximal oxygen uptake, with a unit increase in pulse, x_4, ignoring the values of the other three explanatory variables, which most likely are also changing considering the correlation among the four explanatory variables. In the multiple regression model, −.009 represents a decrease of .009 liters per minute in maximal oxygen uptake, with a unit increase in pulse, x_4, holding the values of the other three explanatory variables constant. Thus, we are considering two groups of subjects having a unit difference in pulse rate, but their age, weight, and time to walk a mile are the same. The difference in the average maximal oxygen uptake between the two groups is .009 liters per minute lower for the group having the larger value for time to walk the mile. ■

model standard deviation

In addition to estimating the intercept and partial slopes, it is important to estimate the **model standard deviation** σ_ε. The residuals, e_i, are defined as before, as the difference between the observed value and the predicted value of y:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik})$$

The sum of squared residuals, SS(Residual), also called SS(Error), is defined exactly as it sounds. Square the prediction errors and sum the squares:

$$\text{SS(Residual)} = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$
$$= \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_i x_{ik})]^2$$

The df for this sum of squares is $n - (k + 1)$. One df is subtracted for the intercept, and 1 df is subtracted for each of the k partial slopes. The mean square residual, MS(Residual), also called MS(Error), is the residual sum of squares divided by $n - (k + 1)$. Finally, the estimate of the model standard deviation s_ε is the square root of MS(Residual).

The estimated model standard deviation s_ε is often referred to as the residual standard deviation. It may also be called "std dev," "standard error of estimate," or "root MSE." If the output is not clear, you can take the square root of MS(Residual) by hand. As always, interpret the standard deviation by the Empirical Rule. About 95% of the prediction errors will be within ± 2 standard deviations of the mean (and the mean error is automatically zero):

$$s_\varepsilon = \sqrt{\text{MS(Residual)}} = \sqrt{\frac{\text{SS(Residual)}}{n - (k + 1)}}$$

EXAMPLE 12.8

The following SPSS computer output is obtained from the data in Example 12.6. Identify SS(Residual) and s_ε in Table 12.12.

TABLE 12.12
SPSS output for Example 12.6

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.763[a]	.582	.547	.29945

a. Predictors: (Constant), pulse, age, wgt, time

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.106	4	1.527	17.024	.000[a]
	Residual	4.394	49	.090		
	Total	10.500	53			

a. Predictors: (Constant), pulse, age, wgt, time
b. Dependent variable: y

Solution In Table 12.12, SPSS labels the table containing the needed information as ANOVA. In this table, SS(Residual) = 4.394 with df = 49. Recall that this data set had $n = 54$ observations and $k = 4$ explanatory variables. Therefore, we confirm the value from the table by computing Residual df $= n - (k + 1) = 54 - (4 + 1) = 49$. Just above the ANOVA table, the value .29945 is given in the column headed by "Std. Error of the Estimate." This is the value of s_ε. We can confirm this value by computing

$$s_\varepsilon = \sqrt{\text{SS(Residual)}/\text{df}} = \sqrt{4.394/49} = .29945 \blacksquare$$

12.4 Inferences in Multiple Regression

We make inferences about any of the parameters in the general linear model (and hence in multiple regression) as we did for β_0 and β_1 in the linear regression model, $y = \beta_0 + \beta_1 x + \varepsilon$.

Before we do this, however, we must introduce the *coefficient of determination*.

coefficient of determination

The **coefficient of determination,** R^2, is defined and interpreted very much like the r^2 value in Chapter 11. (The customary notation is R^2 for multiple regression and r^2 for simple linear regression.) As in Chapter 11, we define the coefficient of determination as the proportion of the variation in the responses, y, that is explained by the model relating y to $x_1, x_2, \cdots, x_k$. For example, if we have the multiple regression model with three x-values, and $R^2_{y \cdot x_1 x_2 x_3} = .736$, then we can account for 73.6% of the variability of the y-values by using the model relating y to x_1, x_2, and x_3. Formally,

$$R^2_{y \cdot x_1 \cdots x_k} = \frac{\text{SS(Total)} - \text{SS(Residual)}}{\text{SS(Total)}}$$

where

$$\text{SS(Total)} = \sum (y_i - \bar{y})^2$$

EXAMPLE 12.9

Referring to the data in Example 12.8, locate the value of $R^2_{y \cdot x_1, x_2, x_3 x_4}$. Using the sum of squares in the ANOVA table, confirm this value.

Solution The required value is listed under R Square, .582 or 58.2%. From the ANOVA table, we have

$$\text{SS(Regression)} = 6.106 \quad \text{SS(Residual)} = 4.394 \quad \text{SS(Total)} = 10.500$$

From these values, we can compute

$$R^2_{y \cdot x_1, x_2, x_3 x_4} = \frac{(10.500 - 4.394)}{10.500} = .582 \ \blacksquare$$

There is no general relation between the multiple R^2 from a multiple regression equation and the individual coefficients of determination $r^2_{yx_1}, r^2_{yx_2}, \ldots, r^2_{yx_k}$ other than that multiple R^2 must be at least as big as any of the individual r^2 values. If all the independent variables are themselves perfectly uncorrelated with each other, then multiple R^2 is just the sum of the individual r^2 values. Equivalently, if all the xs are uncorrelated with each other, SS(Regression) for the all-predictors model is equal to the sum of SS(Regression) values for simple regressions using one x at a time. If the xs are correlated, it is much more difficult to break apart the overall predictive value of $x_1, x_2, \ldots, x_k$ as measured by $R^2_{y \cdot x_1 \cdots xk}$ into separate pieces that can be attributable to x_1 alone, to x_2 alone, . . . , to x_k alone.

collinearity

When the independent variables are themselves correlated, **collinearity** (sometimes called *multicollinearity*) is present. In multiple regression, we are trying to separate out the predictive value of several predictors. When the predictors are highly correlated, this task is very difficult. For example, suppose that we try to explain variation in regional housing sales over time, using gross domestic product (GDP) and national disposable income (DI) as two of the predictors. DI has been almost exactly a fraction of GDP, so the correlation of these two predictors will be

extremely high. Now, is variation in housing sales attributable more to variation in GDP or to variation in DI? Good luck taking those two apart! It is very likely that either predictor alone will explain variation in housing sales almost as well as both together.

Collinearity is usually present to some degree in a multiple regression study. It is a small problem for slightly correlated xs but a more severe one for highly correlated xs. Thus, if collinearity occurs in a regression study—and it usually does to some degree—it is not easy to break apart the overall $R^2_{y \cdot x_1 x_2 \cdots x_k}$ into separate components associated with each x variable. The correlated xs often account for overlapping pieces of the variability in y, so that often, but not inevitably,

$$R^2_{y \cdot x_1 x_2 \cdots x_k} < r^2_{yx1} + r^2_{yx2} + \cdots + r^2_{yxk}$$

sequential sums of squares

Many statistical computer programs will report **sequential sums of squares.** These SS are *incremental* contributions to SS(Regression) when the independent variables enter the regression model in the order you specify to the program. Sequential sums of squares depend heavily on the particular order in which the independent variables enter the model. Again, the trouble is collinearity. For example, if all variables in a regression study are strongly and positively correlated (as often happens in economic data), whichever independent variable happens to be entered first typically accounts for most of the explainable variation in y and the remaining variables add little to the sequential SS. The explanatory power of any x given all the other xs (which is sometimes called the *unique predictive value* of that x) is small. When the data exhibit severe collinearity, separating out the predictive value of the various independent variables is very difficult indeed.

EXAMPLE 12.10

For the data in Example 12.6, interpret the sequential sums of squares (Type I SS) in the following SAS output for the model in which the explanatory variables were entered in the following order: x_1, x_2, x_3, x_4. Would the sequential sums of squares change if we changed the order in which the explanatory variables were entered in the model as x_3, x_1, x_2, x_4?

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	5.58767	1.02985	5.43	<.0001	216.00000
wgt	1	0.01291	0.00283	4.57	<.0001	1.80280
age	1	-0.08300	0.03484	-2.38	0.0211	0.69733
time	1	-0.15817	0.02658	-5.95	<.0001	2.42053
pulse	1	-0.00911	0.00251	-3.64	0.0007	1.18558

Solution The Type I SS column contains the sequential sum of squares. The variable *wgt* by itself accounts for 1.80280 of the total variation in y, maximal oxygen uptake. Adding the variable *age* to a model already containing *wgt* accounts for another 0.69733 of the variation in y. Adding the variable *time* to a model already containing both *wgt* and *age* accounts for another 2.42053 of the variation in y. Finally, adding the variable *pulse* to a model already containing the other three explanatory variables accounts for another 1.18558 of the variation in y. The following SAS output was for a model in which the explanatory variables were entered as x_3, x_1, x_2, x_4.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	5.58767	1.02985	5.43	<.0001	216.00000
time	1	-0.15817	0.02658	-5.95	<.0001	2.68399
wgt	1	0.01291	0.00283	4.57	<.0001	1.70696
age	1	-0.08300	0.03484	-2.38	0.0211	0.52971
pulse	1	-0.00911	0.00251	-3.64	0.0007	1.18558

We can observe that the sequential sums of squares are different for three of the four variables. Now, the variable *time* by itself accounts for 2.68399 of the total variation in y. Adding the variable *wgt* to a model already containing *time* accounts for another 1.70696 of the variation in y. Adding the variable *age* to a model already containing both *time* and *wgt* accounts for another .52971 of the variation in y. The sum of squares for *pulse* remains the same for both models because it is the last variable entered. Recall that in Example 12.7, we computed the correlations among x_1, x_2, x_3, and x_4. The six correlations ranged from $-.255$ to .116. This results in a change in the sequential sums of squares but not too large a change because the four explanatory variables are only weakly correlated. ■

The ideas of Section 12.4 involve point estimation of the regression coefficients and the standard deviation σ_ε. Because these estimates are based on sample data, they will be in error to some extent, and a researcher should allow for that error in interpreting the model. We now present tests about the partial slope parameters in a multiple regression model.

First, we examine a test of an overall null hypothesis about the partial slopes ($\beta_1, \beta_2, \ldots, \beta_k$) in the multiple regression model. According to this hypothesis—H_0: $\beta_1 = \beta_2 = \cdots \beta_k = 0$—none of the variables included in the multiple regression has any predictive value at all. This is the "nullest" of null hypotheses; it says that all those carefully chosen predictors are absolutely useless. The research hypothesis is a very general one—namely, H_a: At least one $\beta_j \neq 0$. This merely says that there is some predictive value somewhere in the set of predictors.

The test statistic is similar to the F statistic of Chapter 11. To state the test, we first define the sum of squares attributable to the regression of y on the variables $x_1, x_2, \ldots, x_k$. We designate this sum of squares as SS(Regression); it is also called SS(Model) or the explained sum of squares. It is the sum of squared differences between the predicted values and the mean y-value.

DEFINITION 12.2

$$\begin{aligned}
\textbf{SS(Regression)} &= \sum(\hat{y}_i - \bar{y})^2 \\
\text{SS(Total)} &= \sum(y_i - \bar{y})^2 \\
&= \text{SS(Regression)} + \text{SS(Residual)} \\
\text{SS(Regression)} &= \text{SS(Total)} - \text{SS(Residual)} \\
\sum(\hat{y}_i - \bar{y})^2 &= \sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2
\end{aligned}$$

Unlike SS(Total) and SS(Residual), we don't interpret SS(Regression) in terms of prediction error. Rather, it measures the extent to which the predictions $\hat{y}_i$ vary. If SS(Regression) = 0, the predicted y-values $(\hat{y})$ are all the same. In such a case, information about the xs is useless in predicting y. If SS(Regression) is large relative to SS(Residual), the indication is that there is real predictive value in the independent variables $x_1, x_2, \ldots, x_k$. We state the test statistic in terms of mean squares rather than sums of squares. As always, a mean square is a sum of squares divided by the appropriate df.

F Test of H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_a: At least one $\beta \neq 0$.

T.S.: $$F = \frac{\text{SS(Regression)}/k}{\text{SS(Residual)}/[n - (k + 1)]} = \frac{\text{MS(Regression)}}{\text{MS(Residual)}}$$

R.R.: With $df_1 = k$ and $df_2 = n - (k + 1)$, reject H_0 if $F > F_\alpha$.

Check assumptions and draw conclusions.

EXAMPLE 12.11

The following SAS output is provided for fitting the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ to the maximal oxygen uptake data of Example 12.6.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6.10624	1.52656	17.02	<.0001
Error	49	4.39376	0.08967		
Corrected Total	53	10.50000			

Root MSE	0.29945	R-Square	0.5815
Dependent Mean	2.00000	Adj R-Sq	0.5474
Coeff Var	14.97236		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.58767	1.02985	5.43	<.0001
x1	1	0.01291	0.00283	4.57	<.0001
x2	1	-0.08300	0.03484	-2.38	0.0211
x3	1	-0.15817	0.02658	-5.95	<.0001
x4	1	-0.00911	0.00251	-3.64	0.0007

Use this information to answer the following questions.

a. Locate SS(Regression).
b. Locate the F statistic.
c. Is there substantial evidence that the four independent variables x_1, x_2, x_3, and x_4 as a group have at least some predictive power? That is, does the evidence support the contention that at least one of the β_js is not zero?

Solution

a. SS(Regression) is shown in the Analysis of Variance table as SS(Model) with a value of 6.10624.

b. The MS(Regression) is given as MS(Model) = 1.52656, which is just SS(Regression)/df = SS(Model)/df = 6.10624/4. MS(Residual) is given as MS(Error) = .08967, which is just SS(Residual)/df = SS(Error)/df = 4.39376/49 = .08967.

The F statistic is given as 17.02, which is computed as follows

$$F = \frac{\text{MS(Regression)}}{\text{MS(Residual)}} = \frac{1.52656}{.08967} = 17.02$$

c. For $df_1 = 4$, $df_2 = 49$, and $\alpha = .01$, the tabled F-value is 3.73. The computed F is 17.02 which is much larger than 3.73. Therefore, there is strong evidence (p-value $< .0001$, much smaller than $\alpha = .01$) in the data to reject the null hypothesis and conclude that the four explanatory variables collectively have at least some predictive value. ■

***F* and R^2**

This ***F*** test may also be stated in terms of $\boldsymbol{R^2}$. Recall that $R^2_{y \cdot x_1 \cdots x_k}$ measures the reduction in squared error for y attributed to how well the xs predict y. Because the regression of y on the xs accounts for a proportion $R^2_{y \cdot x_1 \cdots x_k}$ of the total squared error in y,

$$\text{SS(Regression)} = R^2_{y \cdot x_1 \cdots x_k}\,\text{SS(Total)}$$

The remaining fraction, $1 - R^2$, is incorporated in the residual squared error:

$$\text{SS(Residual)} = (1 - R^2_{y \cdot x_1 \cdots x_k})\,\text{SS(Total)}$$

The overall F test statistic can be rewritten as

$$F = \frac{\text{MS(Regression)}}{\text{MS(Residual)}} = \frac{R^2_{y \cdot x_1 \cdots x_k}/k}{(1 - R^2_{y \cdot x_1 \cdots x_k})/[n - (k + 1)]}$$

This statistic is to be compared with tabulated F-values for $df_1 = k$ and $df_2 = n - (k + 1)$.

EXAMPLE 12.12

A large city bank studies the relation of average account size in each of its branches to per capita income in the corresponding zip code area, number of business accounts, and number of competitive bank branches. The data are analyzed by Statistix, as shown here:

```
CORRELATIONS (PEARSON)

          ACCTSIZE  BUSIN    COMPET
BUSIN    -0.6934
COMPET    0.8196   -0.6527
INCOME    0.4526    0.1492   0.5571

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF ACCTSIZE
```

```
PREDICTOR
VARIABLES     COEFFICIENT     STD ERROR     STUDENT'S T       P       VIF

CONSTANT          0.15085       0.73776          0.20       0.8404
BUSIN            -0.00288     8.894E-04         -3.24       0.0048     5.2
COMPET           -0.00759       0.05810         -0.13       0.8975     7.4
INCOME            0.26528       0.10127          2.62       0.0179     4.3

R-SQUARED            0.7973     RESID. MEAN SQUARE (MSE)    0.03968
ADJUSTED R-SQUARED   0.7615     STANDARD DEVIATION          0.19920

SOURCE        DF       SS          MS         F         P

REGRESSION     3     2.65376     0.88458    22.29    0.0000
RESIDUAL      17     0.67461     0.03968
TOTAL         20     3.32838
```

a. Identify the multiple regression prediction equation.
b. Use the R^2 value shown to test H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. (*Note:* $n = 21$.)

Solution

a. From the output, the multiple regression forecasting equation is

$$\hat{y} = 0.15085 - 0.00288x_1 - 0.00759x_2 + 0.26528x_3$$

b. The test procedure based on R^2 is

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

H_a: At least one β_j differs from zero.

$$\text{T.S.:} \quad F = \frac{R^2_{y \cdot x_1 x_2 x_3}/3}{(1 - R^2_{y \cdot x_1 x_2 x_3})/(21 - 4)} = \frac{.7973/3}{.2027/17} = 22.29$$

R.R.: For $df_1 = 3$ and $df_2 = 17$, the critical .05 value of F is 3.20.

Because the computed F statistic, 22.29, is greater than 3.20, we reject H_0 and conclude that one or more of the x-values has some predictive power. This also follows because the p-value, shown as .0000, is (much) less than .05. Note that the F-value we compute is the same as that shown in the output. ■

Rejection of the null hypothesis of this F test is not an overwhelmingly impressive conclusion. This rejection merely indicates that there is good evidence of *some* degree of predictive value *somewhere* among the independent variables. It does not give any direct indication of how strong the relation is or any indication of which individual independent variables are useful. The next task, therefore, is to make inferences about the individual partial slopes.

To make these inferences, we need the estimated standard error of each partial slope. As always, the standard error for any estimate based on sample data indicates how accurate that estimate should be. These standard errors are computed and shown by most regression computer programs. They depend on three things: the residual standard deviation, the amount of variation in the predictor variable, and the degree of correlation between that predictor and the other predictors. The expression that we present for the standard error is useful in considering the effect of collinearity (correlated independent variables), but it is *not* a particularly good way to do the computation. Let a computer program do the arithmetic.

DEFINITION 12.3

Estimated standard error of $\hat{\beta}_j$ in a multiple regression:

$$s_{\hat{\beta}_j} = s_\varepsilon \sqrt{\frac{1}{\sum (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}}$$

where R_j^2 is the R^2 value obtained by letting x_j be the *dependent* variable in a multiple regression, with all other xs independent variables. Note that s_ε is the residual standard deviation for the multiple regression of y on $x_1, x_2, \ldots, x_k$.

effect of collinearity

As in simple regression, the larger the residual standard deviation, the larger the uncertainty in estimating coefficients. Also, the less variability there is in the predictor, the larger the standard error of the regression coefficient, $s_{\hat{\beta}_j}$. The most important use of the formula for estimated standard error is to illustrate the **effect of collinearity.** If the independent variable x_j is highly collinear with one or more other independent variables, R_j^2 is by definition very large and $1 - R_j^2$ is near zero. Division by a near-zero number yields a very large standard error. Thus, one important effect of severe collinearity is that it results in very large standard errors of partial slopes and, therefore, very inaccurate estimates of those slopes.

variance inflation factor

The term $1/(1 - R_j^2)$ is called the **variance inflation factor** (VIF). It measures how much the variance (square of the standard error) of a coefficient is increased because of collinearity. This factor is printed out by some computer packages and is helpful in assessing how serious the collinearity problem is. If the VIF is 1, there is no collinearity at all. If it is very large, such as 10 or more, collinearity is a serious problem.

A large standard error for any estimated partial slope indicates a large probable error for the estimate. The partial slope $\hat{\beta}_j$ of x_j estimates the effect of increasing x_j by one unit while all other xs remain constant. If x_j is highly collinear with other xs, when x_j increases, the other xs also vary rather than staying constant. Therefore, it is difficult to estimate β_j, and its probable error is large when x_j is severely collinear with other independent variables.

The standard error of each estimated partial slope $\hat{\beta}_j$ is used in a confidence interval and statistical test for $\hat{\beta}_j$. The confidence interval follows the familiar format of estimate plus or minus table value times estimated standard error. The table value is the t table with the error df, $n - (k + 1)$.

DEFINITION 12.4

The confidence interval for β_j is

$$\left(\hat{\beta}_j - t_{\alpha/2} s_{\hat{\beta}_j},\ \hat{\beta}_j + t_{\alpha/2} s_{\hat{\beta}_j}\right)$$

where $t_{\alpha/2}$ cuts off area $\alpha/2$ in the tail of a t distribution with df $= n - (k + 1)$, the error df.

EXAMPLE 12.13

Calculate a 95% confidence interval for β_3, the coefficient associated with the explanatory variable *INCOME* in the three-predictor model for the data of Example 12.12.

Solution The least-squares estimator of β_1 is $\hat{\beta}_1 = .26528$ with standard error $s_{\hat{\beta}_3} = .10127$. The upper .025 percentile of the t distribution with df $= n - (k + 1) = 21 - (3 + 1) = 17$ is 2.110. The 95% confidence interval on β_3 is computed as

$$\hat{\beta}_3 \pm t_{\alpha/2} s_{\hat{\beta}_3} = .26528 \pm (2.110)(.10127) = (.05160, .47896) \blacksquare$$

EXAMPLE 12.14

Locate the estimated partial slope for x_1 and its standard error in the output in Example 12.11. Calculate a 90% confidence interval for β_1.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.58767	1.02985	5.43	<.0001
x1	1	0.01291	0.00283	4.57	<.0001
x2	1	-0.08300	0.03484	-2.38	0.0211
x3	1	-0.15817	0.02658	-5.95	<.0001
x4	1	-0.00911	0.00251	-3.64	0.0007

Solution $\hat{\beta}_1 = .01291$ with standard error .00283. The tabled t-value for $\alpha/2 = .10/2 = .05$ and df $= 54 - (4 + 1) = 49$ is 1.677. The 90% confidence interval is computed as follows

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1} = (.01291 - (1.677)(.00283), .01291 + (1.677)(.00283))$$
$$= (.00816, .01766) \blacksquare$$

interpretation of H_0: $\beta_j = 0$

The usual null hypothesis for inference about β_j is H_0: $\beta_j = 0$. This hypothesis does not assert that x_j has no predictive value by itself. It asserts that it has no *additional* predictive value over and above that contributed by the other independent variables; that is, if all other xs had already been used in a regression model and then x_j was added last, the prediction would not improve. The test of H_0: $\beta_j = 0$ measures whether x_j has any additional (e.g., unique) predictive value. The t test of this H_0 is summarized next.

Summary for Testing β_j

H_0:	1. $\beta_j \leq 0$	H_a: 1. $\beta_j > 0$
	2. $\beta_j \geq 0$	2. $\beta_j < 0$
	3. $\beta_j = 0$	3. $\beta_j \neq 0$
T.S.:	$t = \hat{\beta}_j / s_{\hat{\beta}_j}$	
R.R.:	1. $t > t_\alpha$	
	2. $t < -t_\alpha$	
	3. $\lvert t \rvert > t_{\alpha/2}$	

where t_α cuts off a right-tail area α in the t distribution with df $= n - (k + 1)$.

Check assumptions and draw conclusions.

This test statistic is shown by virtually all multiple regression programs.

EXAMPLE 12.15

Refer to the output given in Example 12.14.

a. Test $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ at the $\alpha = .10$ level.
b. Is the conclusion of the test compatible with the confidence interval?

Solution

a. The test statistic for $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ is

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{.01291}{.00283} = 4.562$$

The .05 upper percentile for the t distribution with df $= 54 - (4 + 1) = 49$ is 1.677. Because the computed value of the test statistic is greater than the tabled value, we conclude there is significant evidence to reject H_0. Thus, x_1 has additional predictive power in the presence of the other three explanatory variables.

b. The 90% confidence interval for β_1 did not include 0, which indicates that $H_0: \beta_1 = 0$ should be rejected at the $\alpha = .10$ level. ■

EXAMPLE 12.16

Refer to Example 12.12. Locate the t statistic for testing $H_0: \beta_3 \leq 0$ versus $H_a: \beta_3 > 0$ in the output given in Example 12.12. Do the data support $H_a: \beta_3 > 0$ at any of the usual values for α?

Solution The t statistics are shown under the heading *STUDENT'S T*. For x_3 (INCOME), the t statistic is 2.62, which is computed as .26528/.10127. With df = 17, the tabled values from the t distribution are 2.576 and 2.898 for $\alpha = .01$ and .005, respectively. Thus, H_0 would be rejected at the $\alpha = .01$ level but not at the $\alpha = .005$ level.

The output lists a p-value under the column heading P. This p-value is for a two-sided alternative hypothesis, $H_a: \beta_3 \neq 0$. The p-value for the one-sided alternative $H_a: \beta_3 > 0$ is given by $p\text{-value} = Pr(t_{17} > 2.62) = 1 - pt(2.62, 17) = .00896 < .01 = \alpha$. ■

The multiple regression F and t tests that we discuss in this chapter test different null hypotheses. It sometimes happens that the F test results in the rejection of $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$, whereas no t test of $H_0: \beta_j = 0$ is significant. In such a case, we can conclude that there is predictive value in the equation as a whole, but we cannot identify the specific variables that have predictive value. Remember that each t test is testing the unique predictive value. Does this variable add predictive value given all the other predictors? When two or more predictor variables are highly correlated among themselves, it often happens that no x_j can be shown to have significant, unique predictive value, even though the xs together have been shown to be useful. If we are trying to predict housing sales based on gross domestic product and disposable income, we probably cannot prove that GDP adds value given DI or that DI adds value given GDP.

12.5 Testing a Subset of Regression Coefficients

***F* test for several β_js**

In the last section, we presented an F test for testing *all* the coefficients in a regression model and a t test for testing *one* coefficient. Another F test of the null hypothesis tests whether *several* of the true coefficients are zero—that is, whether several of the predictors have no value given the others. For example, if we try to predict the

prevailing wage rate in various geographical areas for clerical workers based on the national minimum wage, national inflation rate, population density in the area, and median apartment rental price in the area, we might well want to test whether the variables related to area (density and apartment price) add anything given the national variables.

A null hypothesis for this situation would say that the true coefficients of density and apartment price are zero. According to this null hypothesis, these two independent variables together have no predictive value once minimum wage and inflation are included as predictors.

The idea is to compare the SS(Regression) or R^2 values when density and apartment price are excluded and when they are included in the prediction equation. When they are included, the R^2 is automatically at least as large as the R^2 when they are excluded because we can predict at least as well with more information as with less. Similarly, SS(Regression) will be larger for the complete model. The F test for this null hypothesis tests whether the gain is more than could be expected by chance alone. In general, let k be the total number of predictors, and let g be the number of predictors with coefficients not hypothesized to be zero ($g < k$). Then $k - g$ represents the number of predictors with coefficients that are hypothesized to be zero. The idea is to find SS(Regression) values using all predictors (the **complete model**) and using only the g predictors that do not appear in the null hypothesis (the **reduced model**). Once these have been computed, the test proceeds as outlined next. The notation is easier if we assume that the reduced model contains $\beta_1, \beta_2, \ldots, \beta_g$, so that the variables in the null hypothesis are listed last.

complete and reduced models

F Test of a Subset of Predictors

H_0: $\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$

H_a: H_0 is not true.

T.S.: $$F = \frac{[\text{SS(Regression, complete)} - \text{SS(Regression, reduced)}]/(k - g)}{\text{SS(Residual, complete)}/[n - (k + 1)]}$$

R.R.: $F > F_\alpha$, where F_α cuts off a right tail of area α of the F distribution with $\text{df}_1 = (k - g)$ and $\text{df}_2 = [n - (k + 1)]$.

Check assumptions and draw conclusions.

EXAMPLE 12.17

A state fisheries commission wants to estimate the number of bass caught in a given lake during a season in order to restock the lake with the appropriate number of young fish. The commission could get a fairly accurate assessment of the seasonal catch by extensive "netting sweeps" of the lake before and after a season, but this technique is much too expensive to be done routinely. Therefore, the commission samples a number of lakes and records y, the seasonal catch (thousands of bass per square mile of lake area); x_1, the number of lakeshore residences per square mile of lake area; x_2, the size of the lake in square miles; $x_3 = 1$ if the lake has public access, 0 if not; and x_4, a structure index. (Structures are weed beds, sunken trees, drop-offs, and other living places for bass.) The data are shown in Table 12.13.

The commission is convinced that residences and size are important variables in predicting catch because they both reflect how intensively the lake has been

TABLE 12.13
Bass catch data

Lake	Catch	Residence	Size	Access	Structure
1	3.6	92.2	.21	0	81
2	.8	86.7	.30	0	26
3	2.5	80.2	.31	0	52
4	2.9	87.2	.40	0	64
5	1.4	64.9	.44	0	40
6	.9	90.1	.56	0	22
7	3.2	60.7	.78	0	80
8	2.7	50.9	1.21	0	60
9	2.2	86.1	.34	1	30
10	5.9	90.0	.40	1	90
11	3.3	80.4	.52	1	74
12	2.9	75.0	.66	1	50
13	3.6	70.0	.78	1	61
14	2.4	64.6	.91	1	40
15	.9	50.0	1.10	1	22
16	2.0	50.0	1.24	1	50
17	1.9	51.2	1.47	1	37
18	3.1	40.1	2.21	1	61
19	2.6	45.0	2.46	1	39
20	3.4	50.0	2.80	1	53

fished. However, the commission is uncertain whether access and structure are useful as additional predictor variables. Therefore, two regression models (with all four predictor variables entered linearly) are fitted to the data, the first model with all four variables and the second model without access and structure. The relevant portions of the Minitab output follow.

```
Full Model:

Regression Analysis: catch versus residenc, size, access, structur

The regression equation is
catch = - 2.78 + 0.0268 residenc + 0.504 size + 0.743 access + 0.0511 structur


Predictor     Coef    SE Coef      T      P
Constant   -2.7840     0.8157  -3.41  0.004
residenc  0.026794   0.009141   2.93  0.010
size        0.5035     0.2208   2.28  0.038
access      0.7429     0.2021   3.68  0.002
structur  0.051129   0.004542  11.26  0.000

S = 0.389498  R-Sq = 91.4%  R-Sq(adj) = 89.1%

Analysis of Variance

Source          DF        SS      MS      F      P
Regression       4   24.0624  6.0156  39.65  0.000
Residual Error  15    2.2756  0.1517
Total           19   26.3380
```

```
Reduced Model:

Regression Analysis: catch versus residenc, size

The regression equation is
catch = - 0.87 + 0.0394 residenc + 0.828 size

Predictor      Coef    SE Coef       T      P
Constant    -0.871      2.409   -0.36  0.722
residenc   0.03941    0.02733    1.44  0.168
size        0.8280     0.6372    1.30  0.211

S = 1.17387  R-Sq = 11.1%  R-Sq(adj) = 0.6%

Analysis of Variance

Source          DF        SS      MS      F      P
Regression       2     2.913   1.456   1.06  0.369
Residual Error  17    23.425   1.378
Total           19    26.338
```

a. Write the complete and reduced models.
b. Write the null hypothesis for testing that the omitted variables have no (incremental) predictive value.
c. Perform an F test for this null hypothesis.

Solution

a. The complete and reduced models are, respectively,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

and

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

The corresponding multiple regression least-squares equations based on the sample data are

$$\text{Complete: } \hat{y} = -2.78 + .0268x_1 + .504x_2 + .743x_3 + .0511x_4$$

$$\text{Reduced: } \hat{y} = -.87 + .0394x_1 + .828x_2$$

b. The appropriate null hypothesis of no predictive power for x_3 and x_4 is $H_0: \beta_3 = \beta_4 = 0$.

c. The test statistic for the H_0 of part (b) makes use of SS(Regression, complete) = 24.0624, SS(Regression, reduced) = 2.913, SS(Residual, complete) = 2.2756, $k = 4$, $g = 2$, and $n = 20$:

$$\text{T.S.: } F = \frac{[\text{SS(Regression, complete)} - \text{SS(Regression, reduced)}]/(4-2)}{\text{SS(Residual, complete)}/(20-5)}$$

$$= \frac{(24.0624 - 2.913)/2}{2.2756/(20-5)} = 69.705$$

The tabled value $F_{.01}$ for 2 and 15 df is 6.36. The value of the test statistic is much larger than the tabled value, so we have conclusive evidence that the access and structure variables add predictive value ($p < .0001$). ■

12.6 Forecasting Using Multiple Regression

One of the major uses for multiple regression models is in forecasting a y-value given certain values of the independent x variables. The best-guess forecast is easy; just substitute the specified x-values into the estimated regression equation. In this section, we discuss the relevant standard errors.

As in simple regression, the forecast of y for given x-values can be interpreted two ways. The resulting value can, first, be thought of as the estimate for $E(y)$, the long-run average y-value that results from averaging infinitely many observations of y when the xs have the specified values. The alternative interpretation is that this is the predicted y-value for *one* individual case having the given x-values. The standard errors for both interpretations require matrix algebra ideas that are not required for this text.

Computer programs typically give a standard error for an individual y forecast. This information can also be used to find a standard error for estimating $E(y)$. In most computer outputs, an interval for the mean value is called a *confidence interval;* a forecast interval for an individual value is called a *prediction interval*. The appropriate plus or minus term for constructing an interval can be found by multiplying the standard error by a tabled t-value with df $= n - (k + 1)$. In fact, many computer programs give the plus or minus term directly.

EXAMPLE 12.18

An advertising manager for a manufacturer of prepared cereals wants to develop an equation to predict sales (s) based on advertising expenditures for children's television (c), daytime television (d), and newspapers (n). Data were collected monthly for the previous 30 months (and divided by a price index to control for inflation). A multiple regression is fit, yielding the following Minitab computer output:

```
The regression equation is
s = 0.053 + 0.00562 c + 0.0184 d - 0.00600 n

Predictor         Coef       Stdev   t-ratio       p
Constant        0.0526      0.1374      0.38   0.705
c             0.005618    0.002930      1.92   0.066
d              0.01841     0.01211      1.52   0.141
n            -0.005996    0.004362     -1.37   0.181

s = 0.04736     R-sq = 30.8%     R-sq(adj) = 22.9%

Analysis of Variance

SOURCE      DF         SS         MS        F       p
Regression   3   0.026003   0.008668     3.86   0.021
Error       26   0.058317   0.002243
Total       29   0.084320

SOURCE      DF     SEQ SS
c            1   0.000330
d            1   0.021434
n            1   0.004238

     Fit  Stdev.Fit        95% C.I.          95% P.I.
 0.24686    0.01998  (0.20579,0.28794)   (0.14118,0.35255)
```

a. Write the regression equation.
b. Locate the predicted y-value ($\hat{y}$) when $c = 31$, $d = 5$, and $n = 12$. Locate the lower and upper limits for a 95% confidence interval for $E(y)$ and the upper and lower 95% prediction limits for an individual y-value.

Solution

a. The column labeled Coef yields the equation

$$\hat{y} = .0526 + .005618c + .01841d - .005996n$$

b. The predicted y-value is shown as Fit. As can be verified by substituting $c = 31$, $d = 5$, and $n = 12$ into the equation, the predicted y is .24686. The 95% confidence limits for the mean $E(y)$ are shown in the 95% C.I. part of the output as .20579 to .28794, whereas the wider prediction limits for an individual y-value are .14118 to .35255. ■

extrapolation in multiple regression

The notion of **extrapolation** is more subtle **in multiple regression** than in simple linear regression. In simple regression, extrapolation occurred when we tried to predict y using an x-value that was well beyond the range of the data. In multiple regression, we must be concerned not only about the range of each individual predictor but also about the set of values of several predictors together. It might well be reasonable to use multiple regression to predict the salary of a 30-year-old middle manager or the salary of a middle manager with 25 years of experience, but it would *not* be reasonable to use regression to predict the salary of a 30-year-old middle manager with 25 years of experience! Extrapolation depends not only on the range of each separate x_j predictor used to develop the regression equation but also on the correlations among the x_j values. In the salary prediction example, obviously age and experience will be positively correlated, so the combination of a low age and high amount of experience wouldn't occur in the data. When making forecasts using multiple regression, we must consider not only whether each independent variable value is reasonable by itself but also whether the chosen combination of predictor values is reasonable.

EXAMPLE 12.19

The state fisheries commission hoped to use the data of Example 12.17 to predict the catch at a lake with 8 residences per square mile, a size of .7 square mile, 1 public access, and a structure index of 55 and also for another lake with 48 residences per square mile, a size of 1.0 square mile, 1 public access, and a structure index of 40. The following Minitab output was obtained:

```
Regression Analysis: catch versus residenc, size, access, structur

The regression equation is
catch = - 2.78 + 0.0268 residenc + 0.504 size + 0.743 access + 0.0511 structur


Predictor     Coef    SE Coef      T      P
Constant   -2.7840     0.8157  -3.41  0.004
residenc  0.026794   0.009141   2.93  0.010
size        0.5035     0.2208   2.28  0.038
access      0.7429     0.2021   3.68  0.002
structur  0.051129   0.004542  11.26  0.000

S = 0.389498  R-Sq = 91.4%  R-Sq(adj) = 89.1%
```

```
Predicted Values for New Observations

New
Obs     Fit  SE Fit      95% CI           95% PI
  1  1.3379  0.6119 (0.0337, 2.6420) (-0.2081,2.8838)XX
  2  1.7937  0.2129 (1.3400, 2.2475) ( 0.8476, 2.7398)

XX denotes a point that is an extreme outlier in the predictors.

Values of Predictors for New Observations

New
Obs  residenc   size   access   structur
  1       8.0   0.70     1.00       55.0
  2      48.0   1.00     1.00       40.0
```

Locate the 95% prediction intervals for the two new lakes. Why is the first interval so much wider than the second?

Solution The prediction intervals are given by the respective 95% PI values, $(-.2081, 2.8838)$ for the first lake and $(.8476, 2.7398)$ for the second lake. The first interval carries a warning: *a point that is an extreme outlier in the predictors*. A check of the data for the original 20 lakes reveals no lake had even close to eight residences per square mile. Thus, the prediction for this set of values of the predictors would be an extrapolation well beyond the data used to fit the model. For this case, the problem is with the value for just one of the explanatory variables, residence; the values for the remaining predictor variables are well within the range of the data. ■

12.7 Comparing the Slopes of Several Regression Lines

This topic represents a special case of the general problem of constructing a multiple regression equation for both qualitative and quantitative independent variables. The best way to illustrate this particular problem is by way of an example.

EXAMPLE 12.20

An investigator was interested in comparing the responses of rats to different doses of two drug products (A and B). The study called for a sample of 60 rats of a particular strain to be randomly allocated into two equal groups. The first group of rats was to receive drug A, with 10 rats randomly assigned to each of three doses (5, 10, and 20 mg). Similarly, the 30 rats in group 2 were to receive drug B, with 10 rats randomly assigned to the 5-, 10-, and 20-mg doses. In the study, each rat received its assigned dose, and after a 30-minute observation period, it was scored for signs of anxiety on a 0- to 30-point scale. Assume that a rat's anxiety score is a linear function of the dosage of the drug. Write a model relating a rat's scores to the two independent variables "drug product" and "drug dose." Interpret the βs.

Solution For this experimental situation, we have one qualitative variable (drug product) and one quantitative variable (drug dose). Letting x_1 denote the drug dose, we have the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

where

x_1 = drug dose

$x_2 = 1$ if drug B, $x_2 = 0$ otherwise

The expected value for y in our model is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Substituting $x_2 = 0$ and $x_2 = 1$, respectively, for drugs A and B, we obtain the expected rat anxiety score for a given dose:

Drug A: $E(y) = \beta_0 + \beta_1 x_1$

Drug B: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$

linear regression lines

These two expected values represent **linear regression lines.** The parameters in the model can be interpreted in terms of the slopes and intercepts associated with these regression lines. In particular,

y-intercept

slope

β_0: ***y*-intercept** for drug A regression line

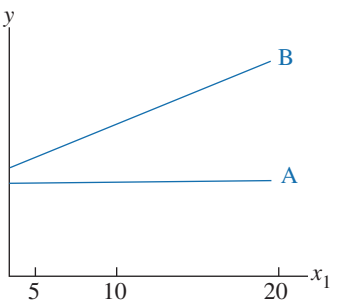
β_1: **slope** of drug A regression line

β_2: difference in y-intercepts of regression lines for drugs B and A

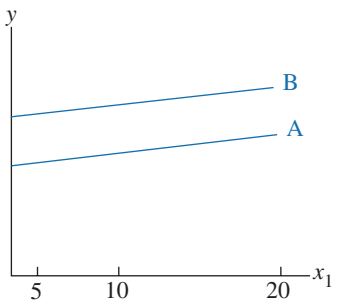
β_3: difference in slopes of regression lines for drugs B and A

Figure 12.4(a) indicates a situation in which $\beta_3 \neq 0$ (that is, there is an interaction between the two variables "drug product" and "drug dose"). Thus, the regression lines are not parallel. Figure 12.4(b) indicates a case in which $\beta_3 = 0$ (no interaction), which results in parallel regression lines.

FIGURE 12.4
Comparing two regression lines



(a) $\beta_3 \neq 0$; interaction is present; intersecting lines



(b) $\beta_3 = 0$; interaction is not present; parallel lines

Indeed, many other experimental situations are possible depending on the signs and magnitudes of the parameters β_0, β_1, β_2, and β_3.

EXAMPLE 12.21

Sample data for the experiment discussed in Example 12.20 are listed in Table 12.14. The response of interest is an anxiety score obtained from trained investigators. Use these data to fit the general linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

TABLE 12.14
Rat anxiety scores

	Drug Dose (mg)					
Drug	**5**		**10**		**20**	
A	15	16	18	16	20	17
	16	15	17	15	19	18
	18	16	18	19	21	21
	13	17	19	18	18	20
	19	15	20	16	19	17
	av = 16		av = 17.6		av = 19.0	
B	16	15	19	18	24	23
	17	15	21	20	25	24
	18	18	22	21	23	22
	17	17	23	22	25	26
	15	16	20	19	25	24
	av = 16.4		av = 20.5		av = 24.1	

Of particular interest to the experimenter is a comparison between the slopes of the regression lines. A difference in slopes would indicate that the drug products have different effects on the anxiety of the rats. Conduct a statistical test of the equality of the two slopes. Use $\alpha = .05$.

Solution Using the complete model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

we obtain a least-squares fit of

$$\hat{y} = 15.30 + .19x_1 - .70x_2 + .30x_1 x_2$$

with SS(Regression, complete) = 442.10 and SS(Residual, complete) = 133.63. (See the computer output that follows.)

The reduced model corresponding to the null hypothesis H_0: $\beta_3 = 0$ (that is, the slopes are the same) is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

for which we obtain

$$\hat{y} = 13.55 + .34x_1 + 2.80x_2$$

and SS(Regression, reduced) = 389.60. The reduction in the sum of squares for error attributed to $x_1 x_2$ is

$$\begin{aligned} \text{SS}_{\text{drop}} &= \text{SS(Regression, complete)} - \text{SS(Regression, reduced)} \\ &= 442.10 - 389.60 = 52.50 \end{aligned}$$

It follows that

$$F = \frac{[\text{SS(Regression, complete)} - \text{SS(Regression, reduced)}]/k - g}{\text{SS(Residual, complete)}/[n - (k + 1)]}$$

$$= \frac{52.50/1}{133.63/56} = 22.00$$

```
REGRESSION ANALYSIS OF ANXIETY TREATMENTS-COMPLETE MODEL


Model: MODEL1
Dependent Variable: SCORE


Analysis of Variance

                          Sum of          Mean
Source          DF       Squares        Square      F Value      Prob>F

Model            3     442.10476     147.36825       61.758      0.0001
Error           56     133.62857       2.38622
C Total         59     575.73333

    Root MSE        1.54474    R-square      0.7679
    Dep Mean       18.93333    Adj R-sq      0.7555
    C.V.            8.15884


Parameter Estimates

                  Parameter      Standard    T for H0:
Variable  DF       Estimate         Error   Parameter=0   Prob > |T|

INTERCEP   1     15.300000    0.59827558         25.573       0.0001
DOSE       1      0.191429    0.04522538          4.233       0.0001
PRODUCT    1     -0.700000    0.84608944         -0.827       0.4116
PRD_DOSE   1      0.300000    0.06395835          4.691       0.0001


               Variable
Variable  DF     Label

INTERCEP   1  Intercept
DOSE       1  DRUG DOSE LEVEL
PRODUCT    1  DRUG PRODUCT
PRD_DOSE   1  PRODUCT TIMES DOSE


REGRESSION ANALYSIS OF ANXIETY TREATMENTS-REDUCED MODEL

Model: MODEL1
Dependent Variable: SCORE


Analysis of Variance

                          Sum of          Mean
Source          DF       Squares        Square      F Value      Prob>F

Model            2     389.60476     194.80238       59.656      0.0001
Error           57     186.12857       3.26541
C Total         59     575.73333

    Root MSE        1.80705    R-square      0.6767
    Dep Mean       18.93333    Adj R-sq      0.6654
    C.V.            9.54425
```

```
Parameter Estimates

                    Parameter      Standard     T for H0:
Variable   DF        Estimate         Error    Parameter=0    Prob > |T|

INTERCEP    1     13.550000    0.54711020         24.766        0.0001
DOSE        1      0.341429    0.03740940          9.127        0.0001
PRODUCT     1      2.800000    0.46657715          6.001        0.0001

                                  Variable
                    Variable   DF     Label

                    INTERCEP    1  Intercept
                    DOSE        1  DRUG DOSE LEVEL
                    PRODUCT     1  DRUG PRODUCT
```

Because the observed value of F exceeds 4.08, the value for $df_1 = 1$, $df_2 = 56$ (actually 40), and $\alpha = .05$ in Appendix Table 8, we reject H_0 and conclude that the slopes for the two groups are different. Note that we could have obtained the same result by testing H_0: $\beta_3 = 0$ using a t test. From the computer output, the t statistic is 4.69, which is significant at the .0001 level. For this type of test, the t statistic and F statistic are related; namely, $t^2 = F$ (here $4.691^2 \approx 22$). ■

The results presented here for comparing the slope of two regression lines can be readily extended to the comparison of three or more regression lines by including additional dummy variables and all possible interaction terms between the quantitative variable x_1 and the dummy variables. Thus, for example, in comparing the slopes of three regression lines, the model would contain the quantitative variable x_1, two dummy variables x_2 and x_3, and two interaction terms x_1x_2 and x_1x_3.

12.8 Logistic Regression

In many research studies, the response variable may be represented as one of two possible values. Thus, the response variable is a binary random variable taking on the values 0 and 1. For example, in a study of a suspected carcinogen, aflatoxin B_1, a number of levels of the compound were fed to test animals. After a period of time, the animals were sacrificed, and the number of animals having liver tumors was recorded. The response variable is $y = 1$ if the animal has a tumor and $y = 0$ if the animal fails to have a tumor. Similarly, a bank wants to determine which customers are most likely to repay their loans. Thus, the bank wants to record a number of independent variables that describe a customer's reliability and then determine whether these variables are related to the binary variable, $y = 1$ if the customer repays the loan and $y = 0$ if the customer fails to repay the loan. A model that relates a binary variable y to explanatory variables will be developed next.

When the response variable y is binary, the distribution of y reduces to a single value, the probability $p = P(y = 1)$. We want to relate p to a linear combination of the independent variables. The difficulty is that p varies between zero and one, whereas linear combinations of the explanatory variables can vary between $-\infty$ and $+\infty$. In Chapter 10, we introduced the transformation of probabilities into an odds ratio. As the probabilities vary between zero and one, the odds ratio varies between zero and infinity. By taking the logarithm of the odds ratio, we will have a transformed variable that will vary between $-\infty$ and $+\infty$ when the probabilities

logistic regression analysis

vary between zero and one. The model often used to study the association between a binary response and a set of explanatory variables is given by **logistic regression analysis.** In this model, the natural logarithm of the odds ratio is related to the explanatory variables by a linear model. We will consider the situation where we have a single independent variable, but this model can be generalized to multiple independent variables. Let $p(x)$ be the probability that y equals 1 when the independent variable equals x. We model the log–odds ratio to a linear model in x, a **simple logistic regression model:**

simple logistic regression model

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

This transformation can be formulated directly in terms of $p(x)$ as

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

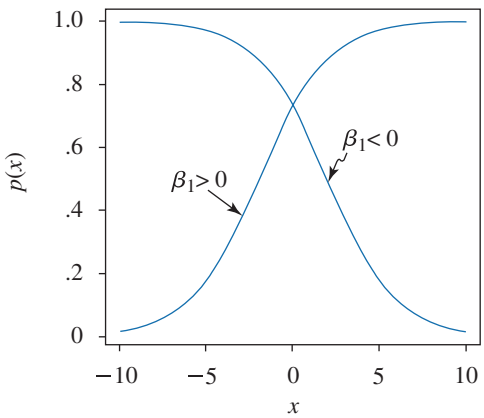
For example, the probability of a tumor being present in an animal exposed to x units of the aflatoxin B_1 would be given by $p(x)$ as expressed by the above equation. The values of β_0 and β_1 would be estimated from the observed data using maximum likelihood estimation.

We can interpret the parameters β_0 and β_1 in the logistic regression model in terms of $p(x)$. The intercept parameter β_0 permits the estimation of the probability of the event associated with $y = 1$ when the independent variable $x = 0$. For example, the probability of a tumor being present when the animal is not exposed to aflatoxin B_1 would correspond to the probability of $y = 1$ when $x = 0$—that is, $p(0)$. The logistic regression model would yield

$$p(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

The slope parameter β_1 measures the degree of association between the probability of the event occurring and the value of the independent variable x. When $\beta_1 = 0$, the probability of the event occurring is not associated with size of the value of x. In our example, the chance of an animal developing a liver tumor would remain constant no matter the amount of aflatoxin B_1 the animal was exposed to. Figure 12.5 displays two simple logistic regression functions. If $\beta_1 > 0$, the probability of the event occurring increases as the value of the independent

FIGURE 12.5
Logistic regression functions



variable increases. If $\beta_1 < 0$, the probability of the event occurring decreases as the value of the independent variable increases.

In the situation where both β_0 and β_1 are zero, the event is as likely to occur as not to occur because

$$p(x) = \frac{e^0}{1 + e^0} = \frac{1}{1 + 1} = \frac{1}{2}$$

This would indicate that the probability of the occurrence of the event indicated by $y = 1$ is not related to the independent variable x. Thus, the model is noninformative in determining the probability of the event's occurrence; there is an equal chance of occurrence or nonoccurence of the event no matter the value of the independent variable.

A second interpretation of the logistic regression model results from using the odds and odds ratio of the event being modeled. For the logistic regression model,

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

and the odds of the event associated with $y = 1$ are

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x} = e^{\beta_0}(e^{\beta_1})^x$$

This exponential relationship provides the following interpretation for the parameter β_1. An increase of one unit in the predictor variable x results in the odds of the specified event being multiplied by e^{β_1}. That is, the odds of the event when the predictor variable equals $x + 1$ equal the odds when the predictor variable has a value of x multiplied by e^{β_1}. Thus, when $\beta_1 = 0$, $e^{\beta_1} = 1$, and, hence, the odds are unchanged when the value of the predictor variable changes. Finally, the odds ratio of the event when the predictor variable has a value $x + 1$ to the event when the predictor variable has a value x is e^{β_1}. This can be seen from the following expression:

$$\frac{p(x + 1)/(1 - p(x + 1))}{p(x)/(1 - p(x))} = \frac{e^{\beta_0 + \beta_1(x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x} = e^{\beta_1}$$

Whether we are using the simple logistic regression model or multiple logistic regression models, the computational techniques used to estimate the model parameters require the use of computer software. We will use an example to illustrate the use of logistic regression models.

EXAMPLE 12.22

A study reported by ***Smith (1967),*** recorded the level of an enzyme, creatinine kinase (CK), for patients who were suspected of having a heart attack. The objective of the study was to assess whether measuring the amount of CK on admission to the hospital was a useful diagnostic indicator of whether patients admitted with a diagnosis of a heart attack had really had a heart attack. The enzyme CK was measured in 360 patients on admission to the hospital. After a period of time, a doctor reviewed the records of these patients to decide which of the 360 patients had actually had a heart attack. The data are given in Table 12.15 with the CK values given as the midpoint of the range of values in each of 13 classes of values.

TABLE 12.15
Heart attack data

CK Value	Number of Patients with Heart Attack	Number of Patients without Heart Attack
20	2	88
60	13	26
100	30	8
140	30	5
180	21	0
220	19	1
260	18	1
300	13	1
340	19	0
380	15	0
420	7	0
460	8	0
500	35	0

The computer output for obtaining the estimated logistic regression curve and 95% confidence intervals on the predicted probabilities of having had a heart attack are given here.

```
                    LOGISTIC REGRESSION ANALYSIS EXAMPLE

           The LOGISTIC Procedure

           Data Set: WORK.LOGREG
           Response Variable (Events): R
           Response Variable (Trials): N
           Number of Observations: 13
           Link Function: Logit

      Response Profile

Ordered  Binary
  Value  Outcome     Count

      1  EVENT         230
      2  NO EVENT      130


Model Fitting Information and Testing Global Null Hypothesis BETA=0

                        Intercept
             Intercept      and
Criterion      Only     Covariates   Chi-Square for Covariates

AIC           472.919      191.773        .
SC            476.806      199.545        .
-2 LOG L      470.919      187.773     283.147 with 1 DF (p=0.0001)
Score            .            .        159.142 with 1 DF (p=0.0001)


          Analysis of Maximum Likelihood Estimates

            Parameter Standard    Wald       Pr >    Standardized
Variable DF  Estimate   Error  Chi-Square Chi-Square   Estimate

INTERCPT 1    -3.0284   0.3670    68.0948     0.0001
CK       1     0.0351  0.00408    73.9842     0.0001     3.100511
```

LOGISTIC REGRESSION ANALYSIS EXAMPLE

OBS	CK	PRED	LCL	UCL
1	20	0.08897	0.05151	0.14937
2	60	0.28453	0.21224	0.36988
3	100	0.61824	0.51935	0.70821
4	140	0.86833	0.78063	0.92436
5	180	0.96410	0.91643	0.98502
6	220	0.99094	0.97067	0.99724
7	260	0.99776	0.99000	0.99950
8	300	0.99945	0.99662	0.99991
9	340	0.99986	0.99886	0.99998
10	380	0.99997	0.99962	1.00000
11	420	0.99999	0.99987	1.00000
12	460	1.00000	0.99996	1.00000
13	500	1.00000	0.99999	1.00000

a. Is CK level significantly related to the probability of a heart attack through the logistic regression model?
b. From the computer output, obtain the estimated coefficients β_0 and β_1.
c. Construct the estimated probability of having had a heart attack as a function of CK level. In particular, estimate this probability for a patient having a CK level of 140.

Solution

a. From the computer output, we obtain, p-value $= .0001$ for testing the hypotheses H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ in the logistic regression model. Thus, there is significant evidence that CK is related to the probability of having had a heart attack.

b. From the computer output, we obtain $\hat{\beta}_0 = -3.0284$ and $\hat{\beta}_1 = .0351$. Note that $\hat{\beta}_1$ is positive. This would indicate that patients having higher levels of CK are associated with a larger probability that a heart attack had occurred. Also, we can conclude that the odds of having had a heart attack for a patient with a CK level of $x + 1$ is $e^{.0351} = 1.036$ times the odds for a patient having a CK level of x.

c. The estimated probability of having had a heart attack as a function of CK level in the patient is given by

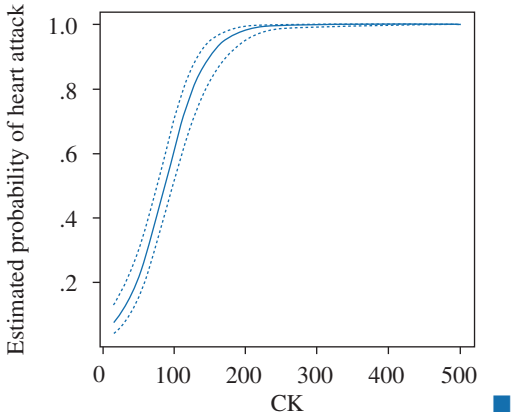
$$p(\widehat{\text{CK}}) = \frac{e^{-3.0284+.0351*\text{CK}}}{1 + e^{-3.0284+.0351*\text{CK}}}$$

We can use this formula to calculate the probability that a patient had experienced a heart attack when the CK level in the patient was 140. This value is given by

$$p(\widehat{\text{CK}}) = \frac{e^{-3.0284+.0351*140}}{1 + e^{-3.0284+.0351*140}} = \frac{e^{1.886}}{1 + e^{1.886}} = .868$$

From the computer printout, we obtain 95% confidence intervals for this probability as .781 to .924. Thus, we are 95% confident that between 78.1% and 92.4% of patients with a CK level of 140 would have had a heart attack. The estimated probabilities of a heart attack along with 95% confidence intervals on these probabilities are plotted in Figure 12.6. We note that the estimated probability of having had a heart attack increases very rapidly with increasing CK levels in the patients. This would indicate that CK levels are a useful indicator of whether a patient has had a heart attack.

FIGURE 12.6
Estimated probability of heart attack with 95% confidence limits



The logistic regression model can be generalized to incorporate k predictor variables. These predictors can be quantitative, qualitative, or a mixture of quantitative and qualitative variables. Let $x_1, x_2, \ldots, x_k$ be the k predictors of the binary response variable y. The logistic regression model given previously generalizes to the following model with x denoting the vector of k predictor variables. The model is given by

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The interpretation of the β_is in this model is similar to the interpretation given to the parameters in the multiple linear regression model. The parameter β_i is related to the effect of the predictor x_i on the log odds ratio that $y = 1$, with the values of the other $k - 1$ predictors held constant. That is, $\exp(\beta_i)$ is the multiplicative effect on the odds of the event occurring for a one-unit increase in the value of the predictor x_i while holding the values of the other $k - 1$ predictors constant.

For example, suppose the values of $x_2, x_3, \ldots, x_k$ are held constant at the values $x_2 = x_{20}, x_3 = x_{30}, \ldots, x_k = x_{k0}$ while the value of x_1 is changed from x_{10} to $x_{10} + 1$. The ratio of the odds that $y = 1$ when $x_2 = (x_{10} + 1, x_{20}, \ldots, x_{k0})$ and when $x_1 = (x_{10}, x_{20}, \ldots, x_{k0})$ is given by

$$\frac{p(x_2)/(1-p(x_2))}{p(x_1)/(1-p(x_1))} = \frac{e^{\beta_0+\beta_1(x_{10}+1)+\beta_2 x_{20}+\cdots+\beta_k x_{k0}}}{e^{\beta_0+\beta_1 x_{10}+\beta_2 x_{20}+\cdots+\beta_k x_{k0}}} = e^{\beta_1} \Rightarrow$$

$$\frac{p(x_2)}{1-p(x_2)} = e^{\beta_1}\frac{p(x_1)}{1-p(x_1)}$$

EXAMPLE 12.23

The following example from **A Handbook of Statistical Analyses *Using SAS (Der and Everitt, 2002)*** will illustrate a logistic regression model with two predictor variables. A study was conducted to examine the extent to which red blood cells settle out of suspension in blood plasma; erythrocyte sedimentation rate (ESR) is related to two proteins that are present in blood plasma. Individuals are classified as healthy (ESR < 20) or unhealthy (ESR ≥ 20). The two blood plasma proteins are fibrinogen (x_1) and γ-globulin (x_2) and are measured on each of the patients in units of grams/liter. The researchers wanted to determine the strength of the relationship between the probability of determining a patient was unhealthy (ESR ≥ 20) and

TABLE 12.16
Blood cell data

Fib	2.52	2.46	2.29	3.15	2.88	2.29	2.99	2.38
Gam	38	36	36	36	30	31	36	37
Hth	0	0	0	0	0	0	0	1
Fib	2.56	3.22	2.35	3.53	2.65	2.15	3.32	2.23
Gam	31	38	29	46	46	31	35	37
Hth	0	0	0	1	0	0	0	0
Fib	2.19	2.21	5.06	2.68	2.09	2.54	2.18	2.67
Gam	33	37	37	34	44	28	31	39
Hth	0	0	1	0	1	0	0	0
Fib	3.41	3.15	3.34	2.60	2.28	3.93	2.60	3.34
Gam	37	39	32	38	36	32	41	30
Hth	0	0	1	0	0	1	0	0

the levels of the two plasma proteins. The data are given in Table 12.16 with Hth (1 = unhealthy, 0 = healthy), Fib (level of fibrinogen), and Gam (level of γ-globulin).

a. Are the levels of the two plasma proteins related to the probability that a patient has an unhealthy level of ESR through the logistic regression model?
b. From the computer output, obtain the estimated coefficients β_0, β_1, and β_2.
c. Construct the estimated probability that a patient has an unhealthy level of ESR as a function of the two predictor variables.
d. From the SAS output, obtain 95% confidence interval on the probabilities for the following two sets of values for the predictor variables: (Fib, Gam) = (2.50, 40) and (Fib, Gam) = (4.90, 38).

```
                    The SAS LOGISTIC Procedure

                        Response Profile

                  Ordered                    Total
                    Value     health     Frequency

                        1          1             6
                        2          0            26

                  Probability modeled is health=1.

              Testing Global Null Hypothesis: BETA=0

       Test                 Chi-Square     DF     Pr > ChiSq

       Likelihood Ratio         7.9138      2         0.0191
       Score                    8.2067      2         0.0165
       Wald                     4.7561      2         0.0927

                    The LOGISTIC Procedure

           Analysis of Maximum Likelihood Estimates

                                Standard          Wald
   Parameter   DF   Estimate       Error    Chi-Square   Pr > ChiSq

   Intercept    1   -12.7920      5.7964        4.8704       0.0273
   fib          1     1.9104      0.9710        3.8708       0.0491
   gamma        1     0.1558      0.1195        1.6982       0.1925
```

```
                    Odds Ratio Estimates

                                 Point          95% Wald
                  Effect      Estimate      Confidence  Limits

                  fib            6.756       1.007       45.308
                  gamma          1.169       0.924        1.477

Obs    fib    gamma    health    _LEVEL_     pred       LCL        UCL

 33   2.50      40       -          1       0.14368    0.03637    0.42724
 34   4.90      38       -          1       0.92332    0.21469    0.99812
```

Solution

a. From the computer output, the likelihood ratio chi-square test has a p-value of .0191. The null hypothesis for this test is H_0: $\beta_1 = 0$, $\beta_2 = 0$. The size of the p-value would suggest that the data support the research hypothesis: H_a: $\beta_1 \neq 0$, and/or $\beta_2 \neq 0$. This would indicate that at least one of the two plasma proteins has predictive power in determining the probability that the patient is unhealthy (ESR $\geq$ 20).

b. The maximum likelihood estimates of the model parameters are

$$\hat{\beta}_0 = -12.7920 \quad \hat{\beta}_1 = 1.9104 \quad \hat{\beta}_2 = .1558$$

c. The estimated equation for obtaining the probability that a patient is unhealthy ($y = 1$) is given by the following equation with $x_1 =$ Fib and $x_2 =$ Gam.

$$\hat{p}(x_1, x_2) = \frac{e^{-12.7920+1.9104x_1+.1558x_2}}{1 + e^{-12.7920+1.9104x_1+.1558x_2}}$$

d. From the SAS output, when Fib equals 2.50 and Gam equals 40, the predicted probability that a patient has these levels of Fib and Gam is .14368 with a 95% confidence interval of (.03637, .42724). When Fib equals 4.90 and Gam equals 38, the predicted probability that a patient has these levels of Fib and Gam is .92332 with a 95% confidence interval of (.21469, .99812). ■

12.9 Some Multiple Regression Theory (Optional)

In this section, we use matrix notation to sketch some of the mathematics underlying multiple regression. The focus is on how multiple regression calculations are actually done, whether by hand or by computer. We do not prove most of the results; proofs are available in many specialized texts, such as Sheather (2009).

First, we will provide a few results related to the algebraic operations on vectors and matrices.

DEFINITION 12.5

A **matrix** B of dimension $m \times n$ is an array of mn elements, assigned to m rows and n columns. Matrices are designated as $B = (b_{ij})$, where b_{ij} represents the number placed in the ith row and jth column of B. A matrix is said to be **square** if $m = n$. A matrix is said to be an **identity** matrix (often designated as I) if it is a square matrix with ones on its diagonal and zeros in all other locations. The **zero** matrix (often designated as **0**) is a matrix with all of its entries equal to zero.

EXAMPLE 12.24

Some examples of matrices are 2×3 matrix $B = \begin{bmatrix} 3 & 8 & 5 \\ 8 & 4 & 1 \end{bmatrix}$; a 3×3 identity matrix $I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$; a 3×3 square matrix

$C = \begin{bmatrix} -2 & 6 & 1.2 \\ 4.3 & 1.2 & 5 \\ 7 & 4 & 1.7 \end{bmatrix}$. ■

A matrix consisting of a single column is called an $m \times 1$ **vector.**

EXAMPLE 12.25

$\mathbf{Y} = \begin{bmatrix} 4 \\ 5 \\ 1 \\ 0 \\ 9 \end{bmatrix}$ is a 5×1 vector. $\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ is a 5×1 **unit** vector.

$\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ is a 3×1 **zero** vector. ■

DEFINITION 12.6

Let C and D be two matrices of the **same** dimension. Then the **addition** and **subtraction** of C and D are given by

$$C + D = (c_{ij} + d_{ij}) \qquad C - D = (c_{ij} - d_{ij})$$

The **multiplication** of matrix C having dimension $m \times n$ by matrix D of dimension $n \times k$ results in the $m \times k$ product matrix $M = CD$ given by

$$M = CD = (m_{ij}) \quad \text{with} \quad m_{ij} = \sum_{t=1}^{n} c_{it} d_{tj}$$

Note that the number of columns of C must equal the number of rows of D in order to multiply C by D.

The **transpose** of an $m \times n$ matrix C is the $n \times m$ matrix C' obtained by placing the rows of C into the columns of C'. A square matrix C is symmetric if $C' = C$.

EXAMPLE 12.26

Let $C = \begin{bmatrix} -2 & 6 & 4 \\ 4 & 2 & 1 \end{bmatrix}$; $D = \begin{bmatrix} 3 & 2 & 4 \\ 9 & 5 & -2 \\ 7 & 1 & 8 \end{bmatrix}$; $E = \begin{bmatrix} 4 & -1 & 0 \\ 8 & 6 & 4 \\ 1 & -6 & 7 \end{bmatrix}$.

Obtain the following matrices: $C + D, D + E, D - E, CD, EC$, and E'.

Solution It is not possible to compute $C + D$ because the two matrices have different dimensions.

$$D + E = \begin{bmatrix} 3+4 & 2+(-1) & 4+0 \\ 9+8 & 5+6 & -2+4 \\ 7+1 & 1+(-6) & 8+7 \end{bmatrix} = \begin{bmatrix} 7 & 1 & 4 \\ 17 & 11 & 2 \\ 8 & -5 & 15 \end{bmatrix}$$

$$D - E = \begin{bmatrix} 3-4 & 2-(-1) & 4-0 \\ 9-8 & 5-6 & -2-4 \\ 7-1 & 1-(-6) & 8-7 \end{bmatrix} = \begin{bmatrix} -1 & 3 & 4 \\ 1 & -1 & -6 \\ 6 & 7 & 1 \end{bmatrix}$$

$$CD = \begin{bmatrix} -2\cdot 3 + 6\cdot 9 + 4\cdot 7 & -2\cdot 2 + 6\cdot 5 + 4\cdot 1 & -2\cdot 4 + 6\cdot -2 + 4\cdot 8 \\ 4\cdot 3 + 2\cdot 9 + 1\cdot 7 & 4\cdot 2 + 2\cdot 5 + 1\cdot 1 & 4\cdot 4 + 2\cdot -2 + 1\cdot 8 \end{bmatrix}$$

$$= \begin{bmatrix} 76 & 30 & 12 \\ 37 & 19 & 20 \end{bmatrix}$$

EC can not be computed because the number of columns in E is 3, whereas the number of rows in C is 2.

$$E' = \begin{bmatrix} 4 & 8 & 1 \\ -1 & 6 & -6 \\ 0 & 4 & 7 \end{bmatrix} \blacksquare$$

DEFINITION 12.7

The **determinant** of a 2×2 square matrix $B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$ is the value $|B| = b_{11}b_{22} - b_{12}b_{21}$.

The **determinant** of a 3×3 square matrix $C = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$ is the value

$$|C| = c_{11}(c_{22}c_{33} - c_{32}c_{23}) - c_{21}(c_{12}c_{33} - c_{32}c_{13}) + c_{31}(c_{12}c_{23} - c_{22}c_{13}).$$

The **inverse** of a square matrix B is the matrix B^{-1} with the property that $BB^{-1} = I$ and $B^{-1}B = I$.

rank

Not all square matrices have an inverse. The **rank** of a matrix is defined as the number of linearly independent rows in the matrix. An $m \times m$ square matrix B has an inverse only if the rank of B is m. If the determinant of a matrix is zero, then the inverse will not exist.

The inverses of 2×2 and 3×3 matrices can be displayed explicitly. For larger matrices, a computer software package should be used to obtain the determinant and inverse.

inverse

The **inverse** of a 2×2 square matrix $B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$ is the matrix

$$B^{-1} = \frac{1}{|B|}\begin{bmatrix} b_{22} & -b_{12} \\ -b_{21} & b_{11} \end{bmatrix}$$

The **inverse** of a 3×3 square matrix $C = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$ is the matrix

$$C^{-1} = \frac{1}{|C|}\begin{bmatrix} c_{22}c_{33} - c_{32}c_{23} & c_{32}c_{13} - c_{12}c_{33} & c_{12}c_{23} - c_{22}c_{13} \\ c_{31}c_{23} - c_{21}c_{33} & c_{11}c_{33} - c_{31}c_{13} & c_{21}c_{13} - c_{11}c_{23} \\ c_{21}c_{32} - c_{31}c_{22} & c_{31}c_{12} - c_{11}c_{32} & c_{11}c_{22} - c_{21}c_{12} \end{bmatrix}$$

EXAMPLE 12.27

Let $B = \begin{bmatrix} 7 & 3 \\ 9 & 5 \end{bmatrix}$ and $C = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 8 & -2 \\ 3 & 1 & 8 \end{bmatrix}$.

Display $|B|$, $|C|$, and the matrices B^{-1} and C^{-1}.

Solution $|B| = 7 \cdot 5 - 9 \cdot 3 = 8$ and

$$|C| = 3(8 \cdot 8 - 1 \cdot -2) - 2(2 \cdot 8 - 1 \cdot 4) + 3(2 \cdot -2 - 8 \cdot 4) = 66$$

$$B^{-1} = \frac{1}{8}\begin{bmatrix} 5 & -3 \\ -9 & 7 \end{bmatrix} = \begin{bmatrix} 5/8 & -3/8 \\ -9/8 & 7/8 \end{bmatrix}$$

Note that $BB^{-1} = B^{-1}B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$.

$$C^{-1} = \frac{1}{66}\begin{bmatrix} 8 \cdot 8 - 1 \cdot -2 & 1 \cdot 4 - 2 \cdot 8 & 2 \cdot -2 - 8 \cdot 4 \\ 3 \cdot -2 - 2 \cdot 8 & 3 \cdot 8 - 3 \cdot 4 & 2 \cdot 4 - 3 \cdot -2 \\ 2 \cdot 1 - 3 \cdot 8 & 3 \cdot 2 - 3 \cdot 1 & 3 \cdot 8 - 2 \cdot 2 \end{bmatrix}$$

$$= \frac{1}{66}\begin{bmatrix} 66 & -12 & -36 \\ -22 & 12 & 14 \\ -22 & 3 & 20 \end{bmatrix}$$

$$= \begin{bmatrix} 66/66 & -12/66 & -36/66 \\ -22/66 & 12/66 & 14/66 \\ -22/66 & 3/66 & 20/66 \end{bmatrix}$$

Note that $CC^{-1} = C^{-1}C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I$. ■

The starting point for the use of matrix notation is the multiple regression model itself. Recall that a model relating a response y to a set of independent variables of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

is called the *general linear model*. The least-squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ of the intercept and partial slopes in the general linear model can be obtained using matrices.

Let the $n \times 1$ matrix $\mathbf{Y}$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

be the matrix of observations, and let the $n \times (k + 1)$ matrix $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}$$

be the matrix of settings for the independent variables augmented with a column of 1s. The first row of $\mathbf{X}$ contains a 1 and the settings for the k independent variables for the first observation, y_1. Row 2 contains a 1 and the corresponding settings for the independent variables for the second observation, y_2. Similarly, the other rows contain settings for the remaining observations.

Next, we turn to the least-squares estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ of the intercept and partial slopes in the multiple regression model. Recall that the least-squares principle involves choosing the estimates to minimize the sum of squared residuals. Those familiar with the calculus will see that the solution can be found by differentiating SS(Residual) with respect to $\hat{\beta}_j (j = 0, \dots, k)$ and setting the result to zero. The resulting normal equations, in matrix notation, are

$$(\mathbf{X}'\mathbf{X})\,\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

where

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

is the desired vector of estimated coefficients. Provided that the matrix $\mathbf{X}'\mathbf{X}$ has an inverse (it does as long as no x_j is perfectly collinear with other xs), the solution is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

EXAMPLE 12.28

Suppose that in a given experimental situation

$$\mathbf{Y} = \begin{bmatrix} 25 \\ 19 \\ 33 \\ 23 \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & -2 & 5 \\ 1 & -2 & -5 \\ 1 & 2 & 5 \\ 1 & 2 & -5 \end{bmatrix}$$

Obtain the least-squares estimates for the prediction equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Solution For these data,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 100 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 100 \\ 24 \\ 80 \end{bmatrix}$$

The $\mathbf{X}'\mathbf{X}$ matrix is a diagonal one, so inverting the matrix is easy. The solution is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$= \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/16 & 0 \\ 0 & 0 & 1/100 \end{bmatrix} \begin{bmatrix} 100 \\ 24 \\ 80 \end{bmatrix} = \begin{bmatrix} 25 \\ 1.5 \\ .8 \end{bmatrix}$$

and the prediction equation is

$$\hat{y} = 25 + 1.5x_1 + .8x_2 \blacksquare$$

The hard part of the arithmetic in multiple regression is computing the inverse of $\mathbf{X}'\mathbf{X}$. For the most realistic multiple regression problems, this task takes hours by hand and fractions of a second by computer. This is the major reason why most multiple regression problems are done with computer software.

Once the inverse of the $\mathbf{X}'\mathbf{X}$ matrix is found and the $\hat{\boldsymbol{\beta}}$ vector is calculated, the next task is to compute the residual standard deviation. The hard work is to compute SS(Residual) $= \sum(y_i - \hat{y}_i)^2$, which can be written as SS(Residual) $= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{Y})$.

EXAMPLE 12.29

Compute SS(Residual) for the data of Example 12.28.

Solution $\hat{\boldsymbol{\beta}}$ and $\mathbf{X}'\mathbf{Y}$ were calculated to be $\begin{bmatrix} 25 \\ 1.5 \\ 0.8 \end{bmatrix}$ and $\begin{bmatrix} 100 \\ 24 \\ 80 \end{bmatrix}$, respectively, and

$$\mathbf{Y}'\mathbf{Y} = [25 \quad 19 \quad 33 \quad 23] \begin{bmatrix} 25 \\ 19 \\ 33 \\ 23 \end{bmatrix} = 2{,}604$$

The shortcut formula yields

$$\text{SS(Residual)} = 2{,}604 - [25 \quad 1.5 \quad .8] \begin{bmatrix} 100 \\ 24 \\ 80 \end{bmatrix} = 2{,}604 - 2{,}600 = 4 \blacksquare$$

Similar calculations yield SS(Regression) and SS(Total). Although the formulas for these sums can be expressed artificially in pure matrix notation, they can be expressed more easily in mixed matrix and algebraic notation:

$$\text{SS(Regression)} = \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{Y}) - \frac{(\Sigma y_i)^2}{n}$$

$$\text{SS(Total)} = \mathbf{Y}'\mathbf{Y} - \frac{(\Sigma y_i)^2}{n}$$

EXAMPLE 12.30

Calculate SS(Regression) and SS(Total) for the data of Example 12.28.

Solution $\Sigma y_i = 100$ and $n = 4$. The relevant matrix calculations were performed in the previous example.

$$\text{SS(Regression)} = 2{,}600 - \frac{(100)^2}{4} = 100$$

$$\text{SS(Total)} = 2{,}604 - \frac{(100)^2}{4} = 104$$

Note that SS(Total) = 104 = 100 + 4 = SS(Regression) + SS(Residual). ■

These sum-of-squares calculations are necessary for making inferences based on R^2 using F tests. For inferences about individual coefficients using t tests, the estimated standard errors of the coefficients are necessary. In Section 12.4, we presented a conceptually useful but computationally cumbersome formula for these estimated standard errors. A much easier way of computing them involves only the standard deviation s_ε and the main diagonal elements of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix.

DEFINITION 12.8

The estimated standard error of $\hat{\boldsymbol{\beta}}_j$ is

$$s_{\hat{\beta}_j} = s_\varepsilon\sqrt{v_{jj}}$$

where s_ε is the standard deviation from the regression equation and v_{jj} is the entry in row $j + 1$, column $j + 1$ of $(\mathbf{X}'\mathbf{X})^{-1}$:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} v_{00} & & & \\ & v_{11} & & \\ & & \ddots & \\ & & & v_{kk} \end{bmatrix}$$

Because the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix must be computed to obtain the $\hat{\boldsymbol{\beta}}_j$s, it is a direct calculation to obtain the estimated standard errors.

EXAMPLE 12.31

Calculate the estimated standard errors of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ for the data of Example 12.28.

Solution

$$s_\varepsilon = \sqrt{MSE} = \sqrt{4/1} = 2$$

$$s_{\hat{\beta}_0} = 2\sqrt{1/4} = 1.0,\ s_{\hat{\beta}_1} = 2\sqrt{1/16} = 0.5$$

$$s_{\hat{\beta}_2} = 2\sqrt{1/100} = 0.2$$ ■

12.10 RESEARCH STUDY: Evaluation of the Performance of an Electric Drill

Defining the Problem

There have been numerous reports of homeowners encountering problems with electric drills. The drills would tend to overheat when under strenuous usage. A consumer product testing laboratory has selected a variety of brands of electric drills to determine what types of drills are most and least likely to overheat under specified conditions. After a careful evaluation of the differences in the designs of the drills, the engineers selected three design factors for use in comparing the resistance of the drills to overheating. The design factors were the thickness of the insulation around the motor, the quality of the wire used in the drill's motor, and the size of the vents in the body of the drill.

Collecting the Data

The engineers designed a study taking into account various combinations of the three design factors. There were five levels of the thickness of the insulation, three levels of the quality of the wire used in the motor, and three sizes for the vents in the drill body. Thus, the engineers had potentially 45 ($5 \times 3 \times 3$) uniquely designed drills. However, each of these 45 drills would have differences with respect to other factors that may impact on their performance. Thus, the engineers selected 10 drills from each of the 45 designs. Another factor that may vary the results of the study is the conditions under which each of the drills is tested. The engineers selected two "torture tests" that they felt reasonably represented the types of conditions under which overheating occurred. The 10 drills were then randomly assigned to one of the two torture tests. At the end of the test, the temperature of the drill was recorded. The mean temperature of the 5 drills was the response variable of interest to the engineers. A second response variable was the logarithm of the sample variance of the 5 drills. This response variable measures the degree to which the 5 drills produced a consistent temperature under each of the torture tests. The goal of the study was to determine which combination of the design factors of the drills produced the smallest values of both response variables. Thus, they would obtain a design for a drill having minimum mean temperature and a design that produced drills for which an individual drill was most likely to produce a temperature closest to the mean temperature.

Summarizing the Data

The data consist of the 90 responses under the various designs and tests. The data were presented in Table 12.4 at the beginning of this chapter with the variables of interest given below.

AVTEM: mean temperature for the five drills under a given torture test
LOGV: logarithm of the variance of the temperatures of the five drills
IT: the thickness of the insulation within the drill (IT = 2, 3, 4, 5, or 6)
QW: an assessment of quality of the wire used in the drill motor (QW = 6, 7, or 8)
VS: the size of the vent used in the motor (VS = 10, 11, or 12)
I2 = (IT − mean IT)^2, Q2 = (QW − mean QW)^2, (V2 = (VS − mean VS)^2
TEST: the type of torture test used

The response variables (dependent variables) are AVTEM and LOGV. The explanatory variables (independent variables) are IT, QW, and VS. Quadratic versions of all three variables will also be considered in finding an appropriate model. These variables are denoted as I2, Q2, and V2. We thus have six possible explanatory variables to be used in our model. There are a total of 90 observations in this study. A preliminary summary of the data is given by the scatterplots in Figures 12.7 and 12.8.

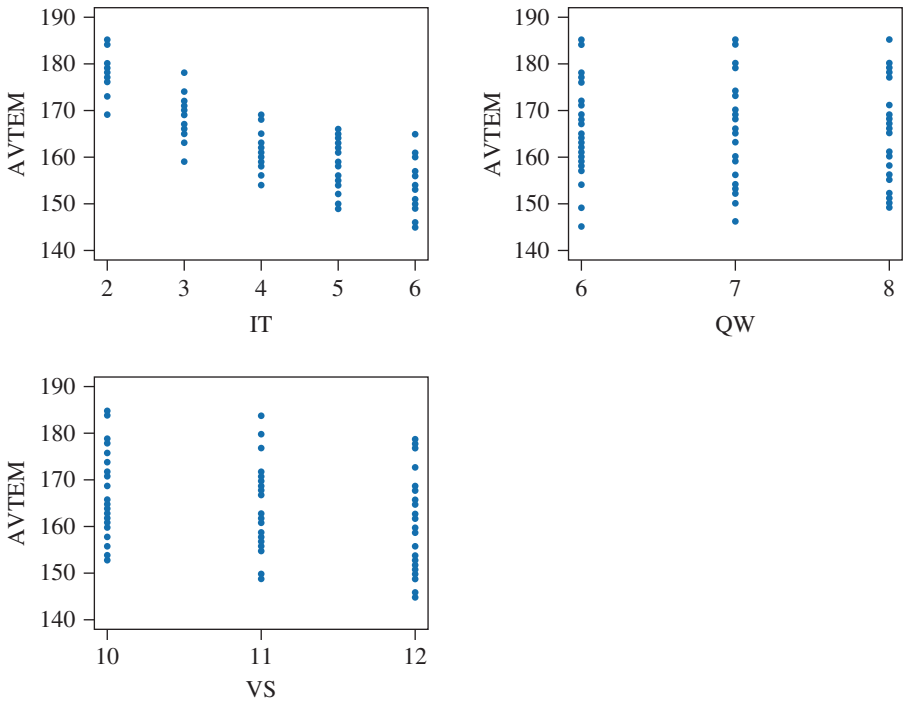


FIGURE 12.7
Scatterplots of IT, QW, and VS versus AVTEM

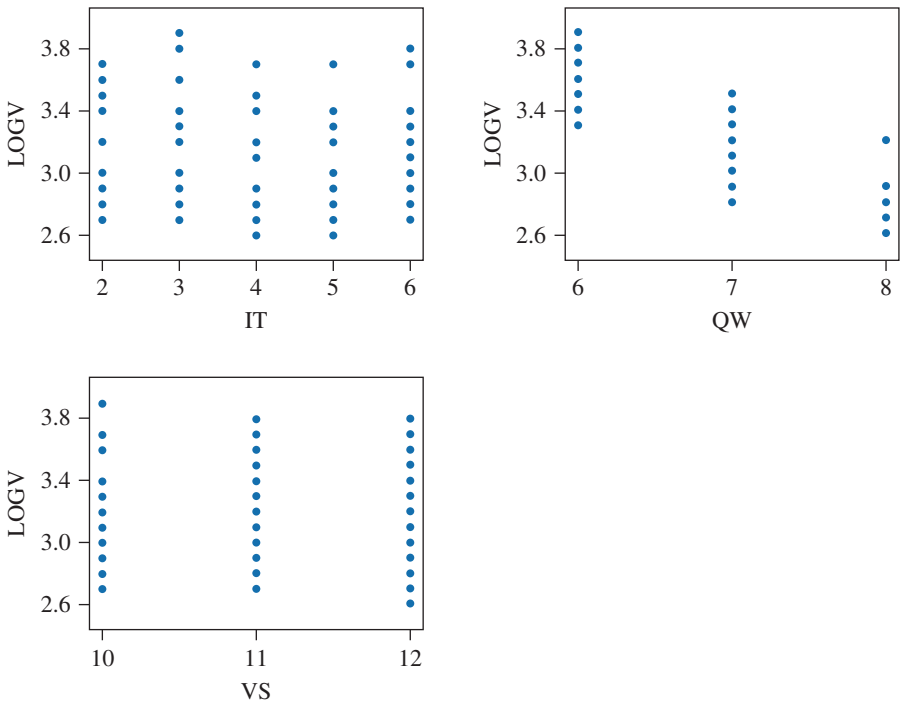


FIGURE 12.8
Scatterplots of IT, QW, and VS versus LOGV

From the scatterplots, the following relationships between the variables are obtained: AVTEM tends to decrease as IT increases—but in a nonlinear fashion. However, AVTEM appears to remain fairly constant with increases in QW and VS. Similarly, LOGV tends to decrease as QW increases—but not at a constant rate. LOGV tends to remain fairly constant with increases in IT and VS.

Analyzing the Data

After examining the scatterplots, the models in Table 12.18 were considered in an attempt to relate AVTEM and LOGV to the explanatory variables.

The goal was to obtain models for AVTEM and LOGV that fit the data well but did not overfit the data. Thus, models were sought that would have a significant fit (small p-value and large R^2 value) without having too many terms in the model. The eight models were programmed for analysis using the SAS software. SAS output is given in Tables 12.18–12.20 using the notation shown in Table 12.17.

TABLE 12.17
Notation for variables in regression models

Variable	Notation	Variable	Notation
IT	x_1	IT*QW	x_7
QW	x_2	IT*VS	x_8
VS	x_3	VS*QW	x_9
I2	x_4	AVTEM	y_1
Q2	x_5	LOGV	y_2
V2	x_6		

TABLE 12.18
Models for describing AVTEM

Models for AVTEM	
Model 1	$\text{AVTEM} = \beta_0 + \beta_1\text{IT} + \beta_2\text{QW} + \beta_3\text{VS} + \varepsilon$
Model 2	$\text{AVTEM} = \beta_0 + \beta_1\text{IT} + \beta_2\text{QW} + \beta_3\text{VS} + \beta_4\text{I2} + \beta_5\text{Q2} + \beta_6\text{V2} + \varepsilon$
Model 3	$\text{AVTEM} = \beta_0 + \beta_1\text{IT} + \beta_2\text{QW} + \beta_3\text{VS} + \beta_4\text{IT} * \text{QW} + \beta_5\text{IT}* \text{VS} + \beta_6\text{QW} * \text{VS} + \varepsilon$
Model 4	$\text{AVTEM} = \beta_0 + \beta_1\text{IT} + \beta_2\text{QW} + \beta_3\text{VS} + \beta_4\text{I2} + \beta_5\text{Q2} + \beta_6\text{V2} + \beta_7\text{IT} * \text{QW} + \beta_8\text{IT} * \text{VS} + \beta_9\text{QW} * \text{VS} + \varepsilon$

```
                          The SAS System
OUTPUT FROM MODELS FOR RELATING AVTEM (y1) TO EXPLANATORY VARIABLES
                    Dependent Variable: y1

MODEL 1:
                      Analysis of Variance

                              Sum of        Mean
Source                 DF     Squares      Square   F Value    Pr > F

Model                   3  7660.94568  2553.64856    131.97    <.0001
Error                  86  1664.17654    19.35089
Corrected Total        89  9325.12222

          Root MSE          4.39896    R-Square    0.8215
          Dependent Mean  164.25556    Adj R-Sq    0.8153


                      Parameter Estimates

                   Parameter   Standard
Variable    DF      Estimate      Error   t Value   Pr > |t|

Intercept    1    234.56106    7.63872     30.71     <.0001
x1           1     -6.15000    0.32788    -18.76     <.0001
x2           1     -0.67445    0.56822     -1.19     0.2385
x3           1     -3.73340    0.56843     -6.57     <.0001
-------------------------------------------------------------
```

```
MODEL 2:
                    Analysis of Variance

                              Sum of        Mean
Source              DF      Squares      Square   F Value    Pr > F

Model                6   7941.21675  1323.53612     79.38    <.0001
Error               83   1383.90547    16.67356
Corrected Total     89   9325.12222

        Root MSE          4.08333    R-Square    0.8516
        Dependent Mean  164.25556    Adj R-Sq    0.8409


                  Parameter Estimates

                 Parameter   Standard
Variable    DF    Estimate      Error   t Value   Pr > |t|

Intercept    1  234.87853    7.16673     32.77     <.0001
x1           1   -6.18215    0.30447    -20.30     <.0001
x2           1   -0.72541    0.52761     -1.37     0.1729
x3           1   -3.81541    0.52812     -7.22     <.0001
x4           1    0.96451    0.24758      3.90     0.0002
x5           1   -0.29207    0.91332     -0.32     0.7499
x6           1   -1.04740    0.91355     -1.15     0.2549
------------------------------------------------------------

MODEL 3:
                    Analysis of Variance

                              Sum of        Mean
Source              DF      Squares      Square   F Value    Pr > F

Model                6   7683.85390  1280.64232     64.76    <.0001
Error               83   1641.26833    19.77432
Corrected Total     89   9325.12222

        Root MSE          4.44683    R-Square    0.8240
        Dependent Mean  164.25556    Adj R-Sq    0.8113


                  Parameter Estimates

                 Parameter   Standard
Variable    DF    Estimate      Error   t Value   Pr > |t|

Intercept    1  214.01181   58.56103      3.65     0.0005
x1           1   -0.53333    5.30316     -0.10     0.9201
x2           1    0.21831    7.91120      0.03     0.9781
x3           1   -2.60819    5.21968     -0.50     0.6186
x7           1   -0.29167    0.40594     -0.72     0.4745
x8           1   -0.32500    0.40594     -0.80     0.4256
x9           1    0.02498    0.70409      0.04     0.9718
------------------------------------------------------------

MODEL 4:
                    Analysis of Variance

                              Sum of        Mean
Source              DF      Squares      Square   F Value    Pr > F

Model                9   7968.16362   885.35151     52.20    <.0001
Error               80   1356.95860    16.96198
Corrected Total     89   9325.12222


        Root MSE          4.11849    R-Square    0.8545
        Dependent Mean  164.25556    Adj R-Sq    0.8381
```

TABLE 12.19
Models for describing LOGV

Models for LOGV	
Model 1	$LOGV = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \varepsilon$
Model 2	$LOGV = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \beta_4 I2 + \beta_5 Q2 + \beta_6 V2 + \varepsilon$
Model 3	$LOGV = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \beta_4 IT * QW + \beta_5 IT * VS + \beta_6 QW * VS + \varepsilon$
Model 4	$LOGV = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \beta_4 I2 + \beta_5 Q2 + \beta_6 V2 + \beta_7 IT * QW + \beta_8 IT * VS + \beta_9 QW * VS + \varepsilon$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	203.41326	54.30065	3.75	0.0003
x1	1	-0.22505	4.91223	-0.05	0.9636
x2	1	1.72599	7.33803	0.24	0.8146
x3	1	-1.82023	4.83905	-0.38	0.7078
x4	1	0.97354	0.25005	3.89	0.0002
x5	1	-0.29587	0.92146	-0.32	0.7490
x6	1	-1.04984	0.92165	-1.14	0.2581
x7	1	-0.34034	0.37617	-0.90	0.3683
x8	1	-0.32500	0.37597	-0.86	0.3899
x9	1	-0.09944	0.65298	-0.15	0.8793

OUTPUT FROM MODELS FOR RELATING LOGV (y2) TO EXPLANATORY VARIABLES
Dependent Variable: y2

MODEL 1:
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9.87413	3.29138	160.33	<.0001
Error	86	1.76543	0.02053		
Corrected Total	89	11.63956			

Root MSE	0.14328	R-Square	0.8483
Dependent Mean	3.19778	Adj R-Sq	0.8430

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.23345	0.24880	25.05	<.0001
x1	1	0.00667	0.01068	0.62	0.5341
x2	1	-0.40568	0.01851	-21.92	<.0001
x3	1	-0.02028	0.01851	-1.10	0.2764

MODEL 2:
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	9.96474	1.66079	82.30	<.0001
Error	83	1.67482	0.02018		
Corrected Total	89	11.63956			

Root MSE	0.14205	R-Square	0.8561
Dependent Mean	3.19778	Adj R-Sq	0.8457

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.25908	0.24932	25.10	<.0001
x1	1	0.00632	0.01059	0.60	0.5525
x2	1	-0.40624	0.01835	-22.13	<.0001
x3	1	-0.02148	0.01837	-1.17	0.2457
x4	1	0.01047	0.00861	1.22	0.2274
x5	1	0.01043	0.03177	0.33	0.7436
x6	1	-0.05300	0.03178	-1.67	0.0991

MODEL 3:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	9.97345	1.66224	82.81	<.0001
Error	83	1.66610	0.02007		
Corrected Total	89	11.63956			

Root MSE	0.14168	R-Square	0.8569
Dependent Mean	3.19778	Adj R-Sq	0.8465

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.95482	1.86582	5.34	<.0001
x1	1	-0.21000	0.16896	-1.24	0.2174
x2	1	-0.81681	0.25206	-3.24	0.0017
x3	1	-0.35718	0.16630	-2.15	0.0347
x7	1	0.00083333	0.01293	0.06	0.9488
x8	1	0.01917	0.01293	1.48	0.1421
x9	1	0.03719	0.02243	1.66	0.1012

MODEL 4:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	10.05889	1.11765	56.57	<.0001
Error	80	1.58066	0.01976		
Corrected Total	89	11.63956			

Root MSE	0.14056	R-Square	0.8642
Dependent Mean	3.19778	Adj R-Sq	0.8489

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.83366	1.85328	5.31	<.0001
x1	1	-0.20686	0.16765	-1.23	0.2209
x2	1	-0.79658	0.25045	-3.18	0.0021
x3	1	-0.34633	0.16516	-2.10	0.0392
x4	1	0.00993	0.00853	1.16	0.2482
x5	1	0.01164	0.03145	0.37	0.7122
x6	1	-0.05187	0.03146	-1.65	0.1031
x7	1	0.00033702	0.01284	0.03	0.9791
x8	1	0.01917	0.01283	1.49	0.1392
x9	1	0.03547	0.02229	1.59	0.1154

The fit of the eight models are summarized in Table 12.21. We will repeat the table of models (Table 12.20) to assist in the evaluation.

TABLE 12.20
Models for describing AVTEM and LOGV

Models for AVTEM	
Model 1	$AVTEM = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \varepsilon$
Model 2	$AVTEM = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \beta_4 I2 + \beta_5 Q2 + \beta_6 V2 + \varepsilon$
Model 3	$AVTEM = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \beta_4 IT * QW + \beta_5 IT * VS + \beta_6 QW * VS + \varepsilon$
Model 4	$AVTEM = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \beta_4 I2 + \beta_5 Q2 + \beta_6 V2 + \beta_7 IT * QW + \beta_8 IT * VS + \beta_9 QW * VS + \varepsilon$
Models for LOGV	
Model 1	$LOGV = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \varepsilon$
Model 2	$LOGV = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \beta_4 I2 + \beta_5 Q2 + \beta_6 V2 + \varepsilon$
Model 3	$LOGV = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \beta_4 IT * QW + \beta_5 IT * VS + \beta_6 QW * VS + \varepsilon$
Model 4	$LOGV = \beta_0 + \beta_1 IT + \beta_2 QW + \beta_3 VS + \beta_4 I2 + \beta_5 Q2 + \beta_6 V2 + \beta_7 IT * QW + \beta_8 IT * VS + \beta_9 QW * VS + \varepsilon$

TABLE 12.21
Model summary information

Model	R^2	Model p-value	p-value for Model Comparisons
Models for AVTEM			
Model 1	.822	<.0001	Model 2 versus Model 1: p-value = .0015
Model 2	.852	<.0001	Model 3 versus Model 1: p-value = .7605
Model 3	.824	<.0001	Model 4 versus Model 3: p-value = .0016
Model 4	.855	<.0001	Model 4 versus Model 2: p-value = .5296
Models for LOGV			
Model 1	.848	< .0001	Model 2 versus Model 1: p-value = .2206
Model 2	.856	< .0001	Model 3 versus Model 1: p-value = .1842
Model 3	.857	< .0001	Model 4 versus Model 3: p-value = .2373
Model 4	.864	< .0001	Model 4 versus Model 2: p-value = .5296

All four models for AVTEM provided a significant (p-value $< .0001$) fit to the data set. The R^2 values for the four models relating AVTEM to the explanatory variables are .822, .852, .824, and .855. There is very little difference in the four values for R^2. Based on the significant fit and the very slight differences in the R^2 values, the most appropriate model would be the model with the fewest independent variables—namely, model 1. Another comparison of the models involves testing whether adding extra terms to model 1 would yield any significant terms in the fitted model. From Table 12.21, only model 2 had added terms over model 1 that were significantly different from 0. That is, the question of examining the addition of terms to model 1 in order to obtain model 2 is equivalent to testing in model 2 the hypotheses

$$H_0: \beta_4 = \beta_5 = \beta_6 = 0 \quad \text{versus} \quad H_a: \text{At least one of } \beta_4, \beta_5, \text{ and } \beta_6 \neq 0.$$

From the SAS output, we obtain the sum of squares model from the two models and compute the value of the F statistic for the full model (model 2) versus the reduced model (model 1):

$$F = \frac{(7{,}941.21675 - 7{,}660.94568)/(6 - 3)}{1{,}383.90547/83} = 5.60 \text{ with df} = 3, 83$$

$$p\text{-value} = P(F_{3,\,83} \geq 5.50) = .0015$$

We thus conclude that model 2 is significantly different in fit than model 1; that is, at least one of β_4, β_5, and β_6 is not equal to 0 in model 2. With p-value = .761, we would conclude that model 3 is not significantly different in fit than model 1; that is, we cannot reject the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ in model 3. With p-value = .530, we would conclude that model 4 is not significantly different in fit than model 2 because we cannot reject the hypothesis that $\beta_7 = \beta_8 = \beta_9 = 0$ in model 4.

Based on the scatterplots and the above test, model 2 would be the most appropriate model. Although model 4 has a slightly larger R^2 value, the F test demonstrates that model 4 is not significantly different from model 2, whereas model 2 is significantly different from model 1. Model 2 includes the variables I2, Q2, and V2, at least one of which appears to significantly improve the fit of the model over model 1. Model 4 is more complex than model 2 but does not appear to provide much improvement in the fit over model 2 (R^2 = .8545 versus .8516).

For the purpose of predicting values of AVTEM, the least-squares estimates produce the following prediction model for AVTEM:

$$\begin{aligned}\text{AVTEM} = {} & 234.879 - 6.182\,\text{IT} - .725\,\text{QW} - 3.815\,\text{VS} + .965\,\text{I2} \\ & -.292\,\text{Q2} - 1.047\,\text{V2}\end{aligned}$$

For the response variable LOGV, all four models provided a significant (p-value < .0001) fit to the data set. The R^2 values for the four models relating LOGV to the explanatory variables are .848, .856, .857, and .864. There is very little difference in the models based on the values for R^2. Based on the significant fit and the very slight differences in the R^2 values, the most appropriate model would be the model with the fewest independent variables—namely, model 1. Another comparison of the models involves testing whether adding extra terms to model 1 would yield any significant terms in the fitted model. From Table 12.21, none of the models provided a significant improvement in fit over model 1. With p-value = .221, we would conclude that model 2 is not significantly different in fit than model 1; that is, we cannot reject the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ in model 2. With p-value = .184, we would conclude that model 3 is not significantly different in fit than model 1; that is, we cannot reject the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ in model 3. With p-value = .237, we would conclude that model 4 is not significantly different in fit than model 3; that is, we cannot reject the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ in model 3. With p-value = .530, we would conclude that model 4 is not significantly different in fit than model 2; that is, we cannot reject the hypothesis that $\beta_7 = \beta_8 = \beta_9 = 0$ in model 4.

Based on the scatterplots, fit statistics, and tests of hypotheses, model 1 would appear to be the most appropriate model. Model 2 and model 3 are not significantly different from model 1. Model 4 is more complex than model 2 but does not provide much improvement in the fit over model 2. Therefore, since the models are not significantly different, the R^2 values are nearly the same, and model 1 is the model containing the fewest independent variables (hence the easiest to understand), I would select model 1. For the purpose of predicting values of LOGV, the least-squares estimates produce the following prediction model LOGV:

$$\text{LOGV} = 6.233 + .00667\,\text{IT} - .406\,\text{QW} - .0203\,\text{VS}$$

12.11 Summary and Key Formulas

This chapter consolidates the material for expressing a response y as a function of one or more independent variables. Multiple regression models (where all the independent variables are quantitative) and models that incorporate information

on qualitative variables were discussed and can be represented in the form of a general linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

After discussing various models and the interpretation of β_s in these models, we presented the normal equations used in obtaining the least-squares estimates $\hat{\beta}$.

A confidence interval and statistical test about an individual parameter β_j were developed using $\hat{\beta}_j$ and the standard error of $\hat{\beta}_j$. We also considered a statistical test about a set of β_s, a confidence interval for $E(y)$ based on a set of xs, and a prediction interval for a given set of xs.

All of these inferences involve a fair to moderate amount of numerical calculation unless statistical software programs are available. Sometimes these calculations can be done by hand if one is familiar with matrix operations (see Section 12.9). However, even these methods become unmanageable as the number of independent variables increases. Thus, the message should be very clear. Inferences about general linear models should be done using available computer software to facilitate the analysis and to minimize computational errors. Our job in these situations is to review and interpret the output.

Aside from a few exercises that will probe your understanding of the mechanics involved with these calculations, most of the exercises in the remainder of this chapter and in the regression problems of the next chapter will make extensive use of computer output.

Here are some reminders about multiple regression concepts:

1. Each regression coefficient in a first-order model (one not containing transformed values, such as squares of a variable or product terms) should be interpreted as a partial slope—the predicted change in a dependent variable when an independent variable is increased by one unit while other variables are held constant.
2. Correlations are important not only between an independent variable and the dependent variable but also between independent variables. Collinearity—correlation between independent variables—implies that regression coefficients will change as variables are added to or deleted from a regression model.
3. The effectiveness of a regression model can be indicated not only by the R^2 value but also by the residual standard deviation.
4. As always, the various statistical tests in a regression model indicate only how strong the evidence is that the apparent pattern is more than random. They don't directly indicate how good a predictive model is. In particular, a large overall F statistic may merely indicate a weak prediction in a large sample.
5. A t test in a multiple regression assesses whether that independent variable adds unique predictive value as a predictor in the model. It is quite possible that several variables may not add a statistically detectable amount of unique predictive value, even though deleting all of them from the model causes a serious drop in predictive value. This is especially true when there is severe collinearity.
6. The variance inflation factor (VIF) is a useful indicator of the overall impact of collinearity in estimating the coefficient of an independent variable. The higher the VIF number, the more serious the impact of collinearity on the accuracy of a slope estimate.
7. Extrapolation in multiple regression can be subtle. Making predictions for a new set of x-values may not be unreasonable when these

values are considered one by one, but the combination of these values may be far outside the range of previous data.

Key Formulas

1. $R^2_{y\cdot x_1\cdots x_k} = \dfrac{\text{SS(Total)} - \text{SS(Residual)}}{\text{SS(Total)}} = \dfrac{\text{SS(Regression)}}{\text{SS(Total)}}$

   where

   $$\text{SS(Total)} = \sum (y_i - \bar{y})^2$$

   $$\text{SS(Regression)} = \sum (\hat{y}_i - \bar{y})^2$$

   $$\text{SS(Residual)} = \sum (y_i - \hat{y}_i)^2$$

2. F test for H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

   $$F = \frac{\text{SS(Regression)}/k}{\text{SS(Residual)}/[n - (k + 1)]}$$

3. $$s_{\hat{\beta}_j} = s_\varepsilon \sqrt{\frac{1}{\sum (x_{ij} - \bar{x})^2 (1 - R^2_{x_j\cdot x_1\cdots x_{j-1}x_{j+1}\cdots x_k})}}$$

   where

   $$s_\varepsilon = \sqrt{\frac{\text{MS(Residual)}}{n - (k + 1)}}$$

4. Confidence interval for β_j

   $$(\hat{\beta}_j - t_{\alpha/2} s_{\hat{\beta}_j}, \hat{\beta}_j + t_{\alpha/2} s_{\hat{\beta}_j})$$

5. Statistical test for β_j

   $$\text{T.S.} \quad t = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

6. Testing a subset of predictors

   $$H_0: \quad \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$$

   $$\text{T.S.} \quad F = \frac{[\text{SS(Regression, complete)} - \text{SS(Regression, reduced)}]/(k - g)}{\text{SS(Residual, complete)}/[n - (k + 1)]}$$

7. Assessing collinearity

   $\text{VIF}_j = 1/(1 - R_j^2)$, where $R_j^2 = R^2_{x_j\cdot x_1\cdots x_{j-1}x_{j+1}\cdots x_k}$

12.12 Exercises

12.2 The General Linear Model

Basic

12.1 An automotive engineer wanted to explain and predict the miles per gallon during city driving, y, for a variety of vehicles, using these explanatory variables: c, the number of cylinders; v, the interior passenger volume; and w, the weight of the vehicle.

a. Write a first-order multiple regression model relating y to c, v, and w.

b. Write a second-order multiple regression model relating y to c, v, w; the squares of c, v, w; and their cross-products.

Basic **12.2** Refer to Exercise 12.1. There are three modes of drive for automobiles: rear-wheel, front-wheel, and all-wheel drive. The engineer wants to relate miles per gallon for the vehicles to the three explanatory variables with a separate model for each mode of drive mechanism.

a. Write a first-order general linear model that allows for different slopes and intercepts for each mode of drive mechanism.

b. In terms of the coefficients of the model in part (a), identify the slopes and intercepts for each of the three modes of drive mechanism.

Basic **12.3** Refer to Exercise 12.2.

a. Write a second-order general linear model that allows for different slopes and intercepts for each mode of drive mechanism.

b. Display the second-order regression equation for each of the three modes of drive mechanism in terms of the coefficients of the model in part (a). *Hint:* A first-order regression model contains terms involving x_i, whereas a second-order regression model involves terms x_i, x_i^2 and $x_i\,x_j$.

Basic **12.4** A cardiologist designs a study to examine factors related to the condition of the heart for patients 50 years of age or older. The design obtains physiological information on hundreds of patients. The cardiologist creates an heart health index, *HHI*, which is an overall assessment of heart health with values ranging from 0 (very poor condition) to 10 (excellent condition). The goal of the study is to obtain a model that will relate (predict) *HHI* to the following explanatory variables: *A*, age; *BMI*, body mass index; *E*, hours of exercise per week; and *SB,* systolic blood pressure reading. Write a first-order regression model relating *HHI* to the four explanatory variables.

Basic **12.5** Refer to Exercise 12.4. The cardiologist decides to include two other variables in the model: an indicator variable for sex, male or female; and an indicator variable for diabetes, yes or no.

a. Write a first-order general linear model that includes sex and diabetes as explanatory variables along with the other four continuous variables.

b. In terms of the coefficients of the model in part (a), display four separate models, one for each combination of the indicator variables sex and diabetes.

Basic **12.6** A researcher employed by the state department of education in a state with a large proportion of students coming from families in which English is not the primary language spoken at home is asked to assess a new program to teach written English to fifth-grade students. She obtains the scores on a statewide language test for 500 fifth-grade students after a year in the new program and for 500 students that were not in the program. Let *S* be the scores on the exam for the 1,000 students. The following model was used to assess the effectiveness of the new program.

$$S = \beta_0 + \beta_1 P + \beta_2 E + \beta_3 P^{*}E + \varepsilon$$

where

$$P = \begin{cases} 1 & \text{if new program} \\ 0 & \text{if old program} \end{cases} \qquad E = \begin{cases} 1 & \text{if English spoken at home} \\ 0 & \text{if English not spoken at home} \end{cases}$$

a. Display the mean score for each of the four groups of students (new program—English spoken at home, old program—English spoken at home, etc.) in terms of the βs in the above model.

b. The researcher wanted to compare the mean scores of students in the new and old programs. She decided it would be necessary to separate the students from homes in which English was spoken from the students from homes in which English was not spoken. For those students whose families did not speak English at home, express the difference in mean scores in terms of the βs in the above model between students who were in the new program and those who were in the old program.

c. For those students whose families spoke English at home, express the difference in mean scores in terms of the βs in the above model between students who were in the new program and those who were in the old program.

d. Explain why it is necessary to include the term $\beta_3 P^{*}E$ in the model.

12.7 Refer to Exercise 12.6. Suppose the researcher wants to determine if the new program is more appropriate for girls than boys. The indicator variable, x_3, was now included in the model, where

$$G = \begin{cases} 1 & \text{if girl} \\ 0 & \text{if boy} \end{cases}$$

A general linear model was used to model S as a function of x_1, x_2, and x_3:

$$S = \beta_0 + \beta_1 P + \beta_2 E + \beta_3 G + \beta_4 P * E + \beta_5 P * G + \beta_6 E * G + \varepsilon$$

a. Using the βs from the above model, express the mean scores for boys and girls enrolled in the new program who lived in homes in which English is not spoken.
b. Express the difference in the mean scores for the new and old programs for girls who lived in homes in which English was not spoken.
c. Express the difference in the mean scores for the new and old programs for boys who lived in homes in which English was spoken.

12.8 Refer to Exercise 12.6. The researcher has decided that a more meaningful evaluation of the new program would use the difference between the score on the exam at the end of the fifth grade and that at the end of the fourth grade. Let S_4 be the score on the exam at the end of the fourth grade and S_5 be the score on the exam at the end of the fifth grade. Initially, the following model was fit to the data set with $D = S_5 - S_4$.

$$\text{Model 1: } D = \beta_0 + \beta_1 P + \beta_2 E + \beta_3 G + \beta_4 P * E + \beta_5 P * G + \beta_6 E * G + \varepsilon$$

It was suggested to the researcher that an improved model would express the fifth-grade score, S_5, as a function of the fourth-grade score, S_4:

$$\text{Model 2: } S_5 = \beta_0 + \beta_1 P + \beta_2 E + \beta_3 G + \beta_4 P * E + \beta_5 P * G + \beta_6 E * G + \beta_7 S_4 + \varepsilon$$

a. Rewrite model 1 to express S_5 as a function of P, E, G, $P * E$, $P * G$, $E * G$, and S_4.
b. Explain why model 2 is a more appropriate model than model 1. *Hint:* Consider your answer to part (a).

12.3 Estimating Multiple Regression Coefficients

Med.

12.9 A pharmaceutical firm would like to obtain information on the relationship between the dose level and potency of a drug product. To do this, each of 15 test tubes is inoculated with a virus culture and incubated for 5 days at 30°C. Three test tubes are randomly assigned to each of the five different dose levels to be investigated (2, 4, 8, 16, and 32 mg). Each tube is injected with only one dose level, and the response of interest (a measure of the protective strength of the product against the virus culture) is obtained. The data are given here.

Dose Level	Response
2	5, 7, 3
4	10, 12, 14
8	15, 17, 18
16	20, 21, 19
32	23, 24, 29

a. Plot the data.
b. Fit linear and quadratic regression models to these data.
c. Which regression equation appears to fit the data better? Why?

12.10 Refer to the data of Exercise 12.9. Often a logarithmic transformation can be used on the dose levels to linearize the response with respect to the independent variable.
a. Obtain the natural logarithms of the five dose levels, ln(dose)
b. Let $x = \ln(\text{dose})$, fit the model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

c. Compare your results to the fitted models in Exercise 12.9. Does the logarithmic transformation provide a better fit than the models in Exercise 12.9?

Med. **12.11** A medical study was conducted to study the relationship between infants' systolic blood pressure and two explanatory variables, weight (kgm) and age (days). The data for 25 infants are shown here.

Infant	Age (Days)	Weight (kgm)	Systolic BP, y
1	3	2.61	80
2	4	2.67	90
3	5	2.98	96
4	6	3.98	102
5	3	2.87	81
6	4	3.41	96
7	5	3.49	99
8	6	4.03	110
9	3	3.41	88
10	4	2.81	90
11	5	3.24	100
12	6	3.75	102
13	3	3.18	86
14	4	3.13	93
15	5	3.98	101
16	6	4.55	103
17	3	3.41	86
18	4	3.35	91
19	5	3.75	100
20	6	3.83	105
21	3	3.18	84
22	4	3.52	91
23	5	3.49	95
24	6	3.81	104
25	6	4.03	107

a. Obtain the estimated regression equation.
b. Obtain the estimated residual standard deviation.
c. Provide an interpretation of $\hat{\beta}_2$, the coefficient of weight.

Bus. **12.12** A regional airline transfers passengers from small airports to a larger regional hub airport. The airline's data analyst was assigned to estimate the revenue (in thousands of dollars) generated by each of the 22 small airports based on two variables: the distance from each airport (in miles) to the hub and the population (in hundreds) of the cities in which each of the 22 airports is located. The data is given in the following table.

Airport	Revenue	Distance	Population	Airport	Revenue	Distance	Population
1	233	233	56	12	267	205	96
2	272	209	74	13	338	214	96
3	253	206	67	14	243	183	73
4	296	232	78	15	252	230	55
5	268	125	73	16	269	238	91
6	296	245	54	17	242	144	64
7	276	213	100	18	233	220	60
8	235	134	98	19	234	170	60
9	253	140	95	20	450	170	240
10	233	165	81	21	340	290	70
11	240	234	52	22	200	340	75

a. Produce three scatterplots: revenue versus distance, revenue versus population, and distance versus population.
b. For the 22 airports, is there a strong correlation between airport distance from the regional hub and city population?
c. Does there appear to a problem with high leverage points? Justify your answer.
d. Fit a first-order regression model relating revenue to distance and population size. Comment on the quality of the fit of the model to the data.
e. Do the two estimated slopes appear to have the appropriate sign? If not, explain why.

Ag. **12.13** A poultry scientist was studying various dietary additives to increase the rate at which chickens gain weight. One of the potential additives was studied by creating a new diet that consisted of a standard basal diet supplemented with varying amounts of the additive (0, 20, 40, 60, 80, and 100 grams). There were 60 chicks available for the study. Each of the six diets was randomly assigned to 10 chicks. At the end of 4 weeks, the feed efficiency ratio, feed consumed (gm) to weight gain (gm), was obtained for the 60 chicks. The data are given here.

Additive	Feed Efficiency Ratio (gm Feed to gm WtGain)
0	1.30, 1.35, 1.44, 1.52, 1.56, 1.61, 1.48, 1.56, 1.45, 1.14
20	2.17, 2.11, 2.08, 2.13, 2.22, 2.29, 2.33, 2.24, 2.16, 2.21
40	2.30, 2.34, 2.20, 2.38, 2.48, 2.44, 2.37, 2.43, 2.37, 2.41
60	2.47, 2.51, 2.79, 2.40, 2.55, 2.67, 2.50, 2.55, 2.60, 2.49
80	3.31, 3.17, 3.24, 3.21, 3.35, 3.38, 3.42, 3.36, 3.25, 3.51
100	4.92, 3.87, 4.81, 4.88, 5.06, 5.09, 4.97, 4.95, 4.59, 4.76

a. In order to explore the relationship between feed efficiency ratio (FER) and feed additive (A), plot the mean FER versus A.
b. What type of regression appears most appropriate?
c. Fit first-order, quadratic, and cubic regression models to the data. Which regression equation provides the best fit to the data? Explain your answer.
d. Is there anything peculiar about any of the data values? Provide an explanation of what may have happened.

Ag. **12.14** Refer to the data of Exercise 12.13. The experiment was also concerned with the effects of high levels of copper in the chick feed. Five of the 10 chicks in each level of the feed additive received 400 ppm of copper, while the remaining five chicks received no copper. The data are given here.

Copper Level	Additive	Feed Efficiency Ratio
0	0	1.30, 1.35, 1.44, 1.52, 1.56
400	0	1.61, 1.48, 1.56, 1.45, 1.14
0	20	2.17, 2.11, 2.08, 2.13, 2.22
400	20	2.29, 2.33, 2.24, 2.16, 2.21
0	40	2.30, 2.34, 2.20, 2.38, 2.48
400	40	2.44, 2.37, 2.43, 2.37, 2.41
0	60	2.47, 2.51, 2.79, 2.40, 2.55
400	60	2.67, 2.50, 2.55, 2.60, 2.49
0	80	3.31, 3.17, 3.24, 3.21, 3.35
400	80	3.38, 3.42, 3.36, 3.25, 3.51
0	100	4.92, 3.87, 4.81, 4.88, 5.06
400	100	5.09, 4.97, 4.95, 4.59, 4.76

Let y be the feed efficiency ratio, x_1 be the amount of the feed additive, and x_2 be the amount of copper placed in the feed. Fit the following two models:

$$\text{Model 1: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$$
$$\text{Model 2: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2 + \varepsilon$$

a. Which of the two models appears to provide the better fit to the data? Justify your answer.
b. Display the predicted equation for the best-fitting model.
c. Explain the meaning of $\hat{\beta}_1$ in the best-fitting model.

12.4 Inferences in Multiple Regression

Bus. **12.15** Refer to Exercise 12.12. Use the fitted regression model to answer the following questions.

a. Can the hypothesis of no overall predictive value of the model be rejected at the the $\alpha = .01$ level?
b. Test the hypothesis that distance to the hub airport is a significant predictor of revenue at the $\alpha = .05$ level.
c. Place a 95% confidence interval of the slope associated with distance to the hub airport.
d. Test the hypothesis that the slope associated with population size is greater than .5 at the $\alpha = .05$ level.

Bus. **12.16** Refer to Exercise 12.12. Fit a second-order regression model to the data.

a. Was there an improvement in the fit of the model compared to the first-order model?
b. Test the hypothesis that distance to the hub airport is a significant predictor of revenue at the $\alpha = .05$ level.
c. Test the hypothesis that population size is a significant predictor of revenue at the $\alpha = .05$ level.

Bus. **12.17** Refer to Exercise 12.12. In the plot of the data, airport 20 had a much larger revenue than any of the other 21 airports.

a. Replot the three scatterplots with the data from airport 20 deleted. Does there appear to be any relationship among revenue and the two explanatory variables in this data set?
b. Fit a first-order regression model relating revenue to distance and population size. Comment on the quality of the fit of the model to the data. Is revenue related to distance from hub and population size once airport 20 is deleted from the data?
c. What conclusions can be inferred from parts (a) and (b) about the importance of plotting the data and not just running models through a software program?

12.18 Refer to Exercise 12.13. Fit a cubic model to the data, and then answer the following questions.

a. Can the hypothesis of no overall predictive value be rejected at the $\alpha = 0.01$ level? Justify your answer.
b. Test the research hypothesis H_0: $\beta_3 = 0$ at the $\alpha = 0.05$ level. Report the p-value of the test.
c. Based on the results of the test in part (b), display the estimated regression model.
d. Plot the data along with the best-fitting estimated regression line.

Med. **12.19** Refer to Exercise 12.11. Fit the following regression model to the data, where y is the systolic blood pressure, A is the age, and W is the weight of the infant.

$$y = \beta_0 + \beta_1 A + \beta_2 W + \beta_3 A^2 + \beta_4 W^2 + \varepsilon$$

a. What are your conclusions about the overall fit of the quadratic model?
b. Conduct a test of the hypothesis that the second-order terms are needed in the model.

c. Does a second- or first-order model appear to be the more appropriate model? Justify your answer.

Med. **12.20** Refer to Exercise 12.19.

a. Test the significance of the four slope parameters in the model using $\alpha = .05$.

b. Are your conclusions from part (a) reasonable considering your results from Exercise 12.19 about the overall fit of the quadratic model?

Med. **12.21** Refer to Exercise 12.19.

a. Provide a 95% confidence interval for the true coefficients associated with age and weight.

b. Interpret the confidence intervals provided in part (a).

Engin. **12.22** A metalworking firm conducts an energy study using multiple regression methods. The dependent variable is y = energy consumption cost per day (in thousands of dollars), and the independent variables are x_1 = tons of metal processed per day, x_2 = average external temperature, x_3 = rated wattage for machinery in use, and $x_4 = x_1x_2$. The data are analyzed by Statistix. Selected output is shown here:

```
CORRELATIONS (PEARSON)

            ENERGY    METAL    METXTEMP  TEMP
METAL       0.6128
METXTEMP    0.4929    0.1094
TEMP        0.4007   -0.0606   0.9831
WATTS       0.5775    0.2239   0.3630    0.3529

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF ENERGY

PREDICTOR
VARIABLES   COEFFICIENT   STD ERROR   STUDENT'S T     P        VIF

CONSTANT       7.20439      17.5322        0.41     0.6855
METAL          1.36291      0.92438        1.47     0.1559     8.8
TEMP           0.30588      1.62104        0.19     0.8522   250.0
WATTS          0.01024      0.00473        2.16     0.0427     1.5
METXTEMP      -0.00277      0.07722       -0.04     0.9717   246.4

R-SQUARED           0.6636    RESID. MEAN SQUARE (MSE)    6.51555
ADJUSTED R-SQUARED  0.5963    STANDARD DEVIATION          2.55255

SOURCE       DF     SS         MS       F        P

REGRESSION    4   257.048    64.2622   9.86   0.0001
RESIDUAL     20   130.311    6.51555
TOTAL        24   387.360

CASES INCLUDED 25  MISSING CASES 0
```

a. Write the estimated model.

b. Summarize the results of the various t tests.

c. Calculate a 95% confidence interval for the coefficient of METXTEMP.

d. What does the VIF column of the output indicate about collinearity problems?

12.5 Testing a Subset of Regression Coefficients

Med. **12.23** Refer to the kinesiology data in Example 12.6. In this example, a first-order model was fit to relate y, maximal oxygen uptake, to the explanatory variables: x_1, weight; x_2, age; x_3, time to walk 1 mile; and x_4, heart rate at the end of a 1-mile walk.

a. Provide the kinesiologist with an interpretation of the fitted model having an R^2 of 58.2%.

b. Fit a quadratic model to the data with the squared values of the four predictors in the model. How much of an increase in R^2 was obtained by this fitting this model?
c. The quadratic model now has eight partial slope coefficients. How many of them are significant at the .05 level?
d. At the .05 level, are the quadratic terms significant taken as a group of four terms?
e. Which of the two model—just first-order terms or first- and second-order terms—would you recommend?

Med. **12.24** Refer to Exercise 12.23. Fit a complete second-order model relating y to x_1, x_2, x_3, and x_4. That is, include both first- and second-order terms in the four variables along with all six cross-product terms, $x_i x_j$.

a. Compare the R^2 values for the three models. Which model appears to provide the best fit?
b. Test at the .05 level whether any of the cross-product terms provide a significant relationship with y, maximal oxygen uptake.
c. Which of the three models would you recommend? Justify your answer.

Ag. **12.25** Refer to the feed efficiency data in Exercise 12.14. The researcher is relating y, feed efficiency, to the explanatory variables: x_1, amount of feed additive; and x_2, amount of copper placed in the feed. Consider the following models:

$$\text{Model 1: } y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$\text{Model 2: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$\text{Model 3: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$$

$$\text{Model 4: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + +\beta_4 x_1 x_2 + \beta_5 x_1^2 x_2\, \varepsilon$$

a. In model 4, which of the coefficients are significantly different from 0 at the .05 level?
b. Do the added terms in model 4 provide a significant gain over model 3 in the fit of the model.
c. Are your conclusions from parts (a) and (b) consistent? Explain in detail.

Med. **12.26** Refer to Exercise 12.25.

a. Compare the R^2 values for the four models. Which model appears to provide the best fit?
b. Test at the .05 level whether the cross-product terms in Model 4 provide a significant relationship with y, feed efficiency.
c. Which of the four models would you recommend? Justify your answer.

Soc. **12.27** An automobile financing company uses a rather complex credit rating system for car loans. The questionnaire requires substantial time to fill out, taking sales staff time and risking alienating the customer. The company decides to see whether three variables (age, monthly family income, and debt payments as a fraction of income) will reproduce the credit score reasonably accurately. Data were obtained on a sample (with no evident biases) of 500 applications. The complicated rating score was calculated and served as the dependent variable in a multiple regression. Some results from JMP are shown.

a. How much of the variation in ratings is accounted for by the three predictors?
b. Use this number to verify the computation of the overall F statistic.
c. Does the F test clearly show that the three independent variables have predictive value for the rating score?

Response: Rating score

Summary of Fit

RSquare	0.979566
RSquare Adj	0.979443
Root Mean Square Error	2.023398
Mean of Response	65.044
Observations (or Sum Wgts)	500

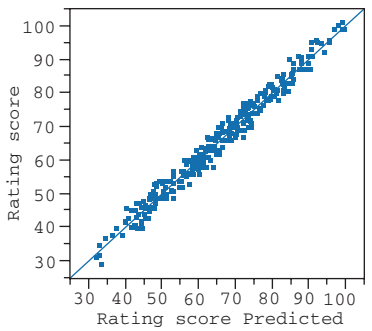
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	54.657197	0.634791	86.10	0.0000
Age	0.0056098	0.011586	0.48	0.6285
Monthly income	0.0100597	0.000157	64.13	0.0000
Debt fraction	-39.95239	0.883684	-45.21	0.0000

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Age	1	1	0.960	0.2344	0.6285
Monthly income	1	1	16835.195	4112.023	0.0000
Debt fraction	1	1	8368.627	2044.05	0.0000

Whole-Model Test



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	97348.339	32449.4	7925.829
Error	496	2030.693	4.1	Prob>F
C Total	499	99379.032		0.0000

12.28 The credit rating data from Exercise 12.27 were reanalyzed, using only the monthly income variable as a predictor. JMP results are shown.

a. By how much has the regression sum of squares been reduced by eliminating age and debt percentage as predictors?

b. Do these variables add statistically significant (at normal α levels) predictive value, once income is given?

Response: Rating score

Summary of Fit

RSquare	0.895261
RSquare Adj	0.895051
Root Mean Square Error	4.571792
Mean of Response	65.044
Observations (or Sum Wgts)	500

Lack of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>[t]
Intercept	30.152827	0.572537	52.67	0.0000
Monthly income	0.0135544	0.000208	65.24	0.0000

Engin. **12.29** A chemical firm tests the yield that results from the presence of varying amounts of two catalysts. Yields are measured for five different amounts of catalyst 1 paired with four different amounts of catalyst 2. A second-order model is fit to approximate the anticipated nonlinear relation. The variables are y = yield, x_1 = amount of catalyst 1, x_2 = amount of catalyst 2, $x_3 = x_1^2$, $x_4 = x_1x_2$, and $x_5 = x_2^2$. Selected output from the regression analysis is shown here.

Multiple Regression Analysis

Dependent variable: Yield

Table of Estimates

	Estimate	Standard Error	t Value	P Value
Constant	50.0195	4.3905	11.39	0.0000
Cat1	6.64357	2.01212	3.30	0.0052
Cat2	7.3145	2.73977	2.67	0.0183
@Cat1Sq	−1.23143	0.301968	−4.08	0.0011
@Cat1Cat2	−0.7724	0.319573	−2.42	0.0299
@Cat2Sq	−1.1755	0.50529	−2.33	0.0355

R-squared = 86.24%
Adjusted R-squared = 81.33%
Standard error of estimation = 2.25973

Analysis of Variance

Source	Sum of Squares	D.F.	Mean Square	F-Ratio	P Value
Model	448.193	5	89.6386	17.55	0.0000
Error	71.489	14	5.10636		
Total (corr.)	519.682	19			

Conditional Sums of Squares

Source	Sum of Squares	D.F.	Mean Square	F-Ratio	P Value
Cat1	286.439	1	286.439	56.09	0.0000
Cat2	19.3688	1	19.3688	3.79	0.0718
@Cat1Sq	84.9193	1	84.9193	16.63	0.0011
@Cat1Cat2	29.8301	1	29.8301	5.84	0.0299
@Cat2Sq	27.636	1	27.636	5.41	0.0355
Model	448.193	5			

```
                    Multiple Regression Analysis

Dependent variable: Yield

                                Table of Estimates

                                Standard        t          P
                 Estimate         Error       Value      Value

Constant         70.31          2.57001       27.36      0.0000
Cat1             -2.676         0.560822      -4.77      0.0002
Cat2             -0.8802        0.70939       -1.24      0.2315

R-squared = 58.85%
Adjusted R-squared = 54.00%
Standard error of estimation = 3.54695

                           Analysis of Variance

                  Sum of                                          P
Source           Squares    D.F.  Mean Square    F-Ratio     Value

Model            305.808      2     152.904        12.15     0.0005
Error            213.874     17      12.5808

Total (corr.)    519.682     19
```

a. Write the estimated complete model.
b. Write the estimated reduced model.
c. Locate the R^2 values for the complete and reduced models.
d. Is there convincing evidence that the addition of the second-order terms improves the predictive ability of the model?

12.6 Forecasting Using Multiple Regression

Med. **12.30** Refer to the data from Exercise 12.11. Recall that a model was fit to relate systolic blood pressure to the age and weight of infants. The researcher wants to be able to predict systolic blood pressure from the fitted model.

a. Provide an estimate for the mean systolic blood pressure for an infant of age 4 days weighing 3 kg.
b. Provide a 95% confidence interval for the mean systolic blood pressure for an infant of age 4 days weighing 3 kg.

Med. **12.31** Refer to Exercise 12.30.

a. Provide an estimate for the mean systolic blood pressure for an infant of age 8 days weighing 5 kg.
b. Provide a 95% confidence interval for the mean systolic blood pressure for an infant of age 8 days weighing 5 kg.

Basic **12.32** The following artificial data are designed to illustrate the effect of correlated and uncorrelated explanatory variables:

y	17	21	26	22	27	25	28	34	29	37	38	38
x	1	1	1	1	2	2	2	2	3	3	3	3
w	1	2	3	4	1	2	3	4	1	2	3	4
v	1	1	2	2	3	3	4	4	5	5	6	6

Here is relevant Minitab output:

```
MTB > Correlation 'y' 'x' 'w' 'v'.

              y        x        w
x         0.856
w         0.402    0.000
v         0.928    0.956    0.262

MTB > Regress 'y' 3 'x' 'w' 'v';
SUBC>   Predict at x 3 w 1 v 6.

The regression equation is
y = 10.0 + 5.00 x + 2.00 w + 1.00 v

s = 2.646     R-sq = 89.5%    R-sq(adj) = 85.6%

    Fit  Stdev.Fit         95% C.I.             95% P.I.
 33.000      4.077   (  23.595,  42.405)  (  21.788,  44.212) XX

X  denotes a row with X values away from the center
XX denotes a row with very extreme X values
```

Locate the 95% prediction interval. Explain why Minitab gave the "very extreme X values" warning.

Med. **12.33** Refer to the kinesiology data in Example 12.6 and the models fit to this data set in Exercises 12.23 and 12.24.

- **a.** Predict the maximal oxygen uptake for a person having a weight of 150 kg, an age of 20 years, a time to walk 1 mile of 17 minutes, and a heart rate of 140 beats per minute using the fitted first-order model. Generate a 95% prediction interval for your prediction.
- **b.** Predict the maximal oxygen uptake for a person having a weight of 150 kg, an age of 20 years, a time to walk 1 mile of 17 minutes, and a heart rate of 140 beats per minute using the fitted second-order model. Generate a 95% prediction interval for your prediction.
- **c.** Predict the maximal oxygen uptake for a person having a weight of 150 kg, an age of 20 years, a time to walk 1 mile of 17 minutes, and a heart rate of 140 beats per minute using the fitted second-order model with cross-product terms. Generate a 95% prediction interval for your prediction.
- **d.** Compare the widths of the three prediction intervals. Did the added complexity of models 2 and 3 provide a substantial reduction in the widths of the intervals?

12.7 Comparing the Slopes of Several Regression Lines

12.34 A psychologist wants to evaluate three therapies for treating people with a gambling addiction. A study is designed to randomly select 25 patients at clinics using each of the three therapies. After the patients had undergone 3 months of inpatient/outpatient treatment, an assessment of each patient's inclination to continue gambling is made, resulting in a gambling inclination score, y, for each patient. The psychologist would like to determine if there is a relationship between the degree to which each patient gambled, as measured by the amount of money the patient had lost gambling the year prior to being admitted to treatment, x, and the gambling score, y. One manner of comparing the difference in the three therapies is to compare the slopes and intercepts of the lines relating y to x.

- **a.** Write a general linear model relating the response, gambling inclination, y, to the explanatory variable, amount of money loss gambling, x, and type of therapy. Make sure to define all variables and parameters in your model.
- **b.** Modify the model of part (a) to reflect that the three therapies have the same slope.

12.35 After sewage is processed through sewage treatment plants, what remains is a dried product called sludge. Sludge contains many minerals that are beneficial to the growth of many farm crops, such as corn, wheat, and barley. Thus, large corporate farms purchase sludge from big cities to use as fertilizer for their crops. However, sludge often contains varying concentrations of heavy metals, which can concentrate in the crops and pose health problems to the people and animals consuming the crops. Therefore, it is important to study the amount of heavy metals absorbed by plants fertilized with sludge. A crop scientist designs the following experiment to study the amount of mercury that may be accumulated in the crops if mercury was contained in sludge. The experiment studied corn, wheat, and barley plants with one of six concentrations of mercury added to the planting soil. There were 90 growth containers used in the experiment with each container having the same soil type. The 18 treatments (three crop types and six mercury concentrations) were randomly assigned to five containers each. At a specified growth stage, the mercury concentration in parts per million (ppm) was determined for the plants in each container. The 90 data values are given here. Note that there are 5 data values for each combination of type of crop and mercury concentration in the soil.

	Type of Crop														
MerCon	Corn					Wheat					Barley				
1	33.3	25.8	24.6	15.1	18.0	17.4	9.2	10.0	25.9	8.6	1.1	23.1	9.6	4.5	8.2
2	31.4	35.7	14.5	40.9	22.9	10.5	34.6	23.4	18.4	24.9	21.2	4.3	9.6	6.4	23.2
3	40.4	35.2	52.1	30.7	46.9	27.1	13.5	30.3	19.3	33.6	30.8	22.0	12.9	3.5	27.9
4	65.6	74.7	77.3	64.2	71.3	50.6	53.9	55.2	48.6	35.2	36.6	34.2	6.8	27.7	39.5
5	94.4	94.9	88.1	100.1	104.8	84.9	77.6	93.3	64.3	74.2	56.7	42.8	49.0	47.9	45.2
6	123.4	158.6	137.3	156.7	133.5	107.5	91.9	87.7	106.2	108.1	70.8	75.7	100.3	64.6	70.1

a. Graph the above data with separate symbols for each crop.
b. Does the relationship between soil mercury content and plant mercury content appear to be linear? Quadratic?
c. Does the relationship between soil mercury content and plant mercury content appear to be the same for all three crops?

12.36 Refer to Exercise 12.35. Fit a single model to the data that will relate x, the soil mercury content, to y, the plant mercury content, with separate intercepts and slopes for the three crops.
a. Does there appear to be a difference in slopes for the three crops?
b. Does there appear to be a difference in intercepts for the three crops?
c. Does a first-order model appear to provide an adequate fit to the data?

12.37 Refer to Exercise 12.36.
a. Write the estimated least-square line for the model without a crop difference.
b. Write the estimated least-square line for the model for each of the three crops.
c. Do the three equations in part (b) appear to be different?

12.38 Refer to Exercise 12.35. Fit a single model to the data that relates y to x and x^2 with separate coefficients for each of the three crops.
a. Does there appear to be a difference in slopes for the three crops?
b. Does there appear to be a difference in intercepts for the three crops?
c. Does a quadratic model appear to provide an adequate fit to the data?

12.39 Refer to Exercise 12.38.
a. Write the estimated least-square quadratic line for the model without a crop difference.
b. Write the estimated least-square quadratic line for the model for each of the three crops.
c. Do the three equations in part (b) appear to be different?

12.8 Logistic Regression

Engin. **12.40** A quality control engineer studied the relationship between years of experience as a system control engineer and the capacity of the engineer to complete within a given time a complex control design including the debugging of all computer programs and control devices. A group of 25 engineers having widely differing amounts of experience (measured in months of experience) was given the same control design project. The results of the study are given in the following table with $y = 1$ if the project was successfully completed in the allocated time and $y = 0$ if the project was not successfully completed.

Months of Experience	Project Success	Months of Experience	Project Success
2	0	15	1
4	0	16	1
5	0	17	0
6	0	19	1
7	0	20	1
8	1	22	0
8	1	23	1
9	0	24	1
10	0	27	1
10	0	30	0
11	1	31	1
12	1	32	1
13	0		

a. Determine whether experience is associated with the probability of completing the task.

b. Compute the probability of successfully completing the task for an engineer having 24 months of experience. Place a 95% confidence interval on your estimate.

12.41 An additive to interior house paint has been recently developed that may greatly increase the ability of the paint to resist staining. An investigation was conducted to determine whether the additive is safe when children are exposed to it. Various amounts of the additive were fed to test animals, and the number of animals developing liver tumors was recorded. The data are given in the following table.

Amount (ppm)	0	10	25	50	100	200
Number of test animals	30	20	20	30	30	30
Number of animals with tumors	0	2	2	7	25	30

a. Determine whether the amount of additive given to the test animals is associated with the probability of a tumor developing in the animals' livers.

b. Compute the probability of a tumor developing in the liver of a test animal exposed to 100 ppm of the additive. Place a 95% confidence interval on your estimate.

12.42 The following example is from the book **Introduction to Regression Modeling *(Abraham and Ledolter, 2006)***. The researchers were examining data on death penalty sentencing in Georgia. For each of 362 death penalty cases, the following information is provided: the outcome (death penalty, yes/no), the race of the victim (white/black), and the aggravation level of the crime. The lowest level (level 1) involved barroom brawls, liquor-induced arguments, and lovers' quarrels. The highest level (level 6) included the most vicious, cruel, cold-blooded, unprovoked crimes.

Aggravation Level	Race of Victim	Death Penalty Yes	Death Penalty No
1	White	2	60
	Black	1	181
2	White	2	15
	Black	1	21
3	White	6	7
	Black	2	9
4	White	9	3
	Black	2	4
5	White	9	0
	Black	4	3
6	White	17	0
	Black	4	0

a. Compute the odds ratio for receiving the death penalty for each of the aggravation levels of the crime.

b. Use a software package to fit the logistic regression model for the variables:

$$y = \begin{cases} 1 & \text{if death} = \text{yes} \\ 0 & \text{if death} = \text{no} \end{cases} \qquad x_1 = \text{aggravation level} \qquad x_2 = \begin{cases} 1 & \text{if black} \\ 0 & \text{if white} \end{cases}$$

c. Is there an association between the severity of the crime and the probability of receiving the death penalty?

d. Is the association between the severity of the crime and the probability of receiving the death penalty different for the two races?

e. Compute the probability of receiving the death penalty for a crime of aggravation level 3 separately for a white and then for a black victim. Place 95% confidence intervals on the two probabilities.

12.9 Some Multiple Regression Theory (Optional)

12.43 Suppose that we have 10 observations on the response variable, y, and two explanatory variables, x_1 and x_2, which are given below in matrix form.

$$\mathbf{Y} = \begin{bmatrix} 25 \\ 31 \\ 26 \\ 38 \\ 18 \\ 27 \\ 29 \\ 17 \\ 35 \\ 21 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1.7 & 10.8 \\ 1 & 6.3 & 9.4 \\ 1 & 6.2 & 7.2 \\ 1 & 6.3 & 8.5 \\ 1 & 10.5 & 9.4 \\ 1 & 1.2 & 5.4 \\ 1 & 1.3 & 3.6 \\ 1 & 5.7 & 10.5 \\ 1 & 4.2 & 8.2 \\ 1 & 6.1 & 7.2 \end{bmatrix}$$

a. Compute $\mathbf{X}'\mathbf{X}$, $(\mathbf{X}'\mathbf{X})^{-1}$, and $\mathbf{X}'\mathbf{Y}$,

b. Compute the least-squares estimators of the prediction equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

12.44 Using the data given in Exercise 12.43, display the $\mathbf{X}$ matrix for the following two prediction models:

a. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$

b. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_1^2 + \hat{\beta}_5 x_2^2$

12.45 Refer to Exercise 12.11. Display the **Y** and **X** matrices for the following two prediction models:

a. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \text{ AGE} + \hat{\beta}_2 \text{ Weight}$

b. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \text{AGE} + \hat{\beta}_2 \text{Weight} + \hat{\beta}_3 \text{AGE}^2 + \hat{\beta}_4 \text{Weight}^2 + \hat{\beta}_5 \text{AGE} \cdot \text{Weight}$

Supplementary Exercises

Bus.

12.46 One of the functions of bank branch offices is to arrange profitable loans to small businesses and individuals. As part of a study of the effectiveness of branch managers, a bank collected data from a sample of branches on current total loan volumes (the dependent variable), total deposits held in accounts opened at that branch, the number of such accounts, the average number of daily transactions, and the number of employees at the branch. Correlations and a scatterplot matrix are shown in the figure.

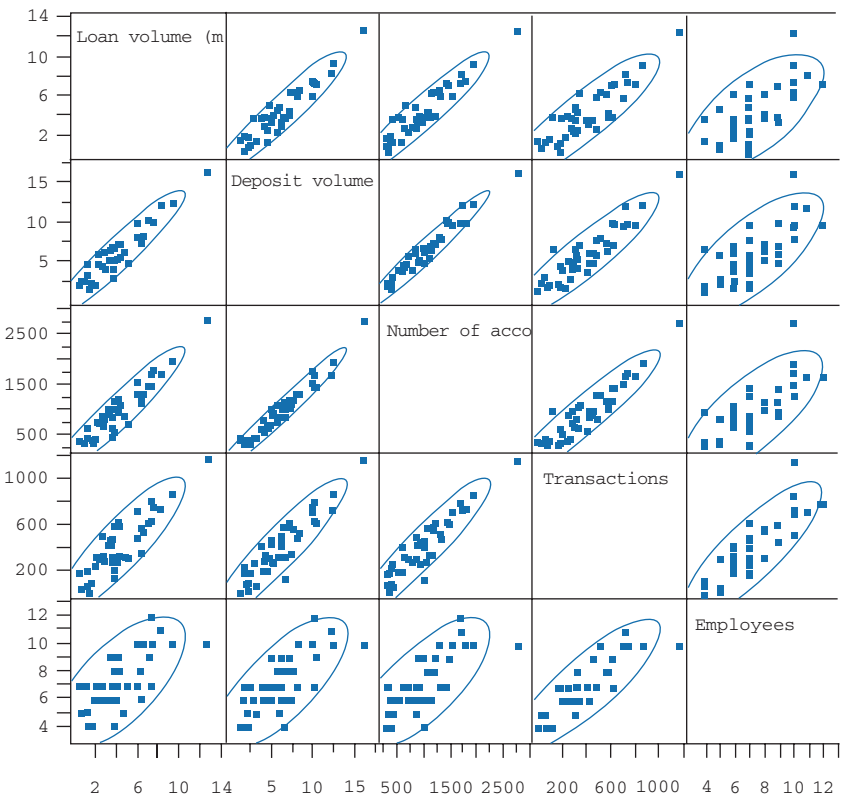
a. Which independent variable is the best predictor of loan volume?

b. Is there a substantial collinearity problem?

c. Do any points seem extremely influential?

Correlations

Variable	Loan volume (millions)	Deposit volume (millions)	Number of accounts	Transactions	Employees
Loan volume (millions)	1.0000	0.9369	0.9403	0.8766	0.6810
Deposit volume (millions)	0.9369	1.0000	0.9755	0.9144	0.7377
Number of accounts	0.9403	0.9755	1.0000	0.9299	0.7487
Transactions	0.8766	0.9144	0.9299	1.0000	0.8463
Employees	0.6810	0.7377	0.7487	0.8463	1.0000



12.47 Refer to Exercise 12.46. A regression model was created for the bank branch office data using JMP. Some of the results are shown here.

a. Use the R^2 value shown to compute an overall F statistic. Is there clear evidence that there is predictive value in the model, using $\alpha = .01$?

b. Which individual predictors have been shown to have unique predictive value, again using $\alpha = .01$?

c. Explain the apparent contradiction between your answers to the first two parts.

Response: Loan volume (millions)

Summary of Fit	
RSquare	0.894477
RSquare Adj	0.883369
Root Mean Square Error	0.870612
Mean of Response	4.383395
Observations(or Sum Wgts)	43

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.2284381	0.6752	0.34	0.7370
Deposit volume (millions)	0.3222099	0.191048	1.69	0.0999
Number of accounts	0.0025812	0.001314	1.96	0.0569
Transactions	0.0010058	0.001878	0.54	0.5954
Employees	-0.119898	0.130721	-0.92	0.3648

12.48 Refer to Exercise 12.46. Another multiple regression model used only deposit volume and number of accounts as independent variables, with results as shown here.

a. Does omitting the transactions and employees variables seriously reduce R^2?

b. Use the R^2 values to test the null hypothesis that the coefficients of transactions and employees are zero. What is your conclusion?

Response: Loan volume (millions)

Summary of Fit	
RSquare	0.892138
RSquare Adj	0.886744
Root Mean Square Error	0.857923
Mean of Response	4.383395
Observations(or Sum Wgts)	43

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.324812	0.290321	-1.12	0.2699
Deposit volume (millions)	0.3227636	0.187509	1.72	0.0929
Number of accounts	0.002684	0.001166	2.30	0.0266

12.49 The following exercise is from *Introduction to Regression Modeling* and refers to data taken from ***Higgins and Koch's, "Variable Selection and Generalized Chi-Square Analysis of Categorical Data Applied to a Large Cross-Sectional Occupational Health Survey" [International Statistical Review (1977) 45:51–62]***. The data were taken from a large survey of workers in the cotton industry. The researchers wanted to study the factors that may be associated with brown lung disease resulting from inhaling particles of cotton, flax, hemp, or jute. The variables are as follows: number of workers suffering from disease (yes); number of workers not suffering from disease (no); dustiness of workplace (1—high; 2—medium; 3—low); race (1—white; 2—other); sex (1—male; 2—female); smoking history (1—smoker; 2—nonsmoker); length of employment in cotton industry (1—less than 10 years; 2—between 10 and 20 years; 3—more than 20 years).

Yes	No	Dust	Race	Sex	Smoking	Employ	Yes	No	Dust	Race	Sex	Smoking	Employ
3	37	1	1	1	1	1	2	8	1	1	1	2	2
0	74	2	1	1	1	1	1	16	2	1	1	2	2
2	258	3	1	1	1	1	0	58	3	1	1	2	2
25	139	1	2	1	1	1	1	9	1	2	1	2	2
0	88	2	2	1	1	1	0	0	2	2	1	2	2
3	242	3	2	1	1	1	0	7	3	2	1	2	2
0	5	1	1	2	1	1	0	0	1	1	2	2	2
1	93	2	1	2	1	1	0	30	2	1	2	2	2
3	180	3	1	2	1	1	1	90	3	1	2	2	2
2	22	1	2	2	1	1	0	0	1	2	2	2	2
2	145	2	2	2	1	1	0	4	2	2	2	2	2
3	260	3	2	2	1	1	0	4	3	2	2	2	2
0	16	1	1	1	2	1	31	77	1	1	1	1	3
0	35	2	1	1	2	1	1	141	2	1	1	1	3
0	134	3	1	1	2	1	12	495	3	1	1	1	3
6	75	1	2	1	2	1	10	31	1	2	1	1	3
1	47	2	2	1	2	1	0	1	2	2	1	1	3
1	122	3	2	1	2	1	0	45	3	2	1	1	3
0	4	1	1	2	2	1	0	1	1	1	2	1	3
1	54	2	1	2	2	1	3	91	2	1	2	1	3
2	169	3	1	2	2	1	3	176	3	1	2	1	3
1	24	1	2	2	2	1	0	1	1	2	2	1	3
3	142	2	2	2	2	1	0	0	2	2	2	1	3
4	301	3	2	2	2	1	0	2	3	2	2	1	3
8	21	1	1	1	1	2	5	47	1	1	1	2	3
1	50	2	1	1	1	2	0	39	2	1	1	2	3
1	187	3	1	1	1	2	3	182	3	1	1	2	3
8	30	1	2	1	1	2	3	15	1	2	1	2	3
0	5	2	2	1	1	2	0	1	2	2	1	2	3
0	33	3	2	1	1	2	0	23	3	2	1	2	3
0	0	1	1	2	1	2	0	2	1	1	2	2	3
1	33	2	1	2	1	2	3	187	2	1	2	2	3
2	94	3	1	2	1	2	2	340	3	1	2	2	3
0	0	1	2	2	1	2	0	0	1	2	2	2	3
0	4	2	2	2	1	2	0	2	2	2	2	2	3
0	3	3	2	2	1	2	0	3	3	2	2	2	3

a. List the five covariates from most likely to least likely to be associated with the probability that a cotton worker has brown lung disease.
b. Do there appear to be any interactions between the covariates?
c. Use a statistical software package to obtain a prediction model using all five covariates.

12.50 Refer to Exercise 12.49. The researchers decide to use the model with all five covariates.
a. Display the estimated probability that a cotton worker will have brown lung disease as a function of the five covariates.
b. Compute the probability that a male white cotton worker who smokes and has worked more than 20 years in a medium-dust workplace will have brown lung disease.
c. Place a 95% confidence interval on your probability from part (b).

Bus. **12.51** A chain of small convenience food stores performs a regression analysis to explain variation in sales volume among 16 stores. The variables in the study are as follows:

Sales:	Average daily sales volume of a store in thousands of dollars
Size:	Floor space in thousands of square feet
Parking:	Number of free parking spaces adjacent to the store
Income:	Estimated per household income of the zip code area of the store

Output from a regression program (StataQuest) is shown here:

```
. regress Sale Size Parking Income

  Source |       SS       df       MS       Number of obs =      16
---------+------------------------------    F(  3,    12) =   15.16
   Model |  27.1296056     3  9.04320188    Prob > F      =  0.0002
Residual |  7.15923792    12   .59660316    R-square      =  0.7912
---------+------------------------------    Adj R-square  =  0.7390
   Total |  34.2888436    15   2.2859229    Root MSE      =   .7724

------------------------------------------------------------------------------
   Sales |      Coef.  Std. Err.     t    P>|T|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Size |   2.547936  1.200827   2.122  0.055  -.0684405   5.164313
 Parking |   .2202793  .1553877   1.418  0.182  -.1182814   .5588401
  Income |   .5893221  .1780576   3.310  0.006   .2013679   .9772763
   _cons |    .872716  1.945615   0.449  0.662  -3.366415   5.111847
------------------------------------------------------------------------------

. correlate  Sales Size Parking Income
(obs=16)

         |   Sales    Size   Parking    Income
---------+------------------------------------
   Sales |  1.0000
    Size |  0.7415   1.0000
 Parking |  0.6568   0.6565   1.0000
  Income |  0.7148   0.4033   0.3241    1.0000
```

a. Write the regression equation. Indicate the standard errors of the coefficients.
b. Carefully interpret each coefficient.
c. Locate R^2 and the residual standard deviation.
d. Is there a severe collinearity problem in this study?

12.52 Summarize the results of the F and t tests for the output of Exercise 12.51.

Ag.

12.53 A producer of various feed additives for cattle conducts a study of the number of days of feedlot time required to bring beef cattle to market weight. Eighteen steers of essentially identical age and weight are purchased and brought to a feedlot. Each steer is fed a diet with a specific combination of protein content, antibiotic concentration, and percentage of feed supplement. The data are as follows:

STEER	1	2	3	4	5	6	7	8	9
PROTEIN	10	10	10	10	10	10	15	15	15
ANTIBIO	1	1	1	2	2	2	1	1	1
SUPPLEM	3	5	7	3	5	7	3	5	7
TIME	88	82	81	82	83	75	80	80	75
STEER	10	11	12	13	14	15	16	17	18
PROTEIN	15	15	15	20	20	20	20	20	20
ANTIBIO	2	2	2	1	1	1	2	2	2
SUPPLEM	3	5	7	3	5	7	3	5	7
TIME	77	76	72	79	74	75	74	70	69

a. Obtain the regression equation relating feedlot time to the three diet variables.
b. Find the value of $\hat{\sigma}_\varepsilon$.
c. Find the R^2 value.
d. How much of a collinearity problem is there with these data?

12.54 Refer to Exercise 12.53.

a. Predict the feedlot time required for a steer fed 15% protein, 1.5% antibiotic concentration, and 5% supplement.

b. Do these values of the independent variables represent a major extrapolation from the data?

c. Give a 95% confidence interval for the mean time predicted in part (a).

12.55 Analyze the data of Exercise 12.53 using a regression model with only protein content as an independent variable.

a. Display the regression equation.

b. Find the R^2 value.

c. Test the null hypothesis that the coefficients of ANTIBIO and SUPPLEM are zero at $\alpha = .05$.

H.R. **12.56** A survey of information systems managers was used to predict the yearly salary of beginning programmer/analysts in a metropolitan area. Managers specified their standard salary for a beginning programmer/analyst, the number of employees in the firm's information processing staff, the firm's gross profit margin in cents per dollar of sales, and the firm's information processing cost as a percentage of total administrative costs. The data are given below for the 68 programmer positions as follows: programmer/analyst, yearly salary, number of employees, profit margin, and information processing cost.

Managers	Programmer/ Analyst Salary	NumEmp	Margin	IPCost	Managers	Programmer/ Analyst Salary	NumEmp	Margin	IPCost
1	29.5	58	19.4	10.14	35	29	38	21.9	6.45
2	29.3	37	17.7	9.18	36	29.2	80	20.9	10.07
3	29.8	135	20.4	6.84	37	28.1	77	14	7.06
4	29.2	69	20.5	7.59	38	27.7	28	19.8	9.7
5	28.9	48	19.1	4.96	39	27.3	30	6.7	3.16
6	31.7	159	23.3	10.52	40	31.3	34	21.4	10.91
7	27.5	42	23.4	8.61	41	27.4	28	16	8.19
8	29.4	37	23.1	10.72	42	29.3	230	14.9	5.7
9	30.4	71	18.57	5.65	43	28.7	121	19.3	6.42
10	27.7	69	16.4	5.46	44	29.7	146	20.9	5.74
11	30.9	121	24.6	7.37	45	29.3	124	17.6	6.13
12	28.9	389	11	7.4	46	28.3	40	16.3	8.86
13	29.7	99	20.9	9.05	47	25.7	130	15.6	4.11
14	30.3	62	23	8.81	48	27.2	60	15.9	6.13
15	31.3	107	15.3	10.94	49	29.2	94	22.6	9.95
16	30	42	18.8	6.84	50	30.2	43	19.6	7.83
17	30	35	21	6.45	51	30.7	111	18.2	6.7
18	28.5	42	10.5	6.06	52	29.4	37	23	11.25
19	29.9	31	19.3	10.2	53	28.4	76	15.5	4.77
20	29.7	78	18	9.6	54	30.1	188	18.9	5.94
21	30.2	132	23.5	7.88	55	28.5	64	12.6	4.81
22	29.7	37	22.4	6.71	56	28.8	185	17.7	8.66
23	29.9	89	22.8	10.04	57	32.4	371	22.3	7.45
24	29	101	21.7	8.39	58	28.4	81	23.1	5.14
25	29.4	60	18	5.24	59	29.7	62	20.9	9.26
26	30.3	48	21.9	9.6	60	27	30	9.8	1.44
27	30.4	75	22.6	11.63	61	28.2	103	22.1	7.98
28	31.1	71	24.5	9.65	62	27.6	29	9.7	6.09
29	29.4	47	24.2	7.94	63	30.7	28	17.1	8.71
30	30.7	39	22.7	9.67	64	28.7	34	16.8	5.11
31	30.2	50	23.1	9.66	65	29.4	279	23.2	6.2
32	30.7	40	16.1	10.31	66	29.9	35	23.4	8.42
33	28.5	102	16.2	6.67	67	31.3	43	18.3	7.52
34	28.5	77	19	7.85	68	28.5	64	12.6	4.81

a. Obtain a multiple regression equation with salary as the dependent variable and the other three variables as predictors. Interpret each of the (partial) slope coefficients.
b. Is there conclusive evidence that the three predictors together have at least some value in predicting salary? Locate a p-value for the appropriate test.
c. Which of the independent variables, if any, have statistically detectable ($\alpha = .05$) predictive value as the last predictor in the equation?

H.R. **12.57** **a.** Compute the coefficient of determination (R^2) for the regression model in Exercise 12.56.
b. Obtain another regression model with number of employees as the only independent variable. Compute the coefficient of determination for this model.
c. Test the null hypothesis that adding profit margin and information processing cost does not yield any additional predictive value given the information about number of employees. Use $\alpha = .10$. What can you conclude from this test?

H.R. **12.58** Obtain correlations for all pairs of predictor variables in Exercise 12.56. Does there seem to be a major collinearity problem in the data?

Gov. **12.59** A government agency pays research contractors a fee to cover overhead costs, over and above the direct costs of a research project. Although overhead costs vary considerably among contracts, they are usually a substantial share of the total contract cost. An agency task force obtained data on overhead costs are a percentage of direct costs, number of employees of the contractor, size of contract as a percentage of the contractor's yearly income, and personnel costs as a percentage of direct costs. The data are given below for the 86 research contractors as follows: contractor, overhead costs as a percentage of direct costs, number of employees, size of contract, and personnel costs as a percentage of direct costs.

Cont.	OverCost	NumEmp	Size	PerCosts	Cont.	OverCost	NumEmp	Size	PerCosts
1	66.4	293	2.14	69	26	78.1	194	0.85	64
2	73.7	117	1.15	61	27	64.0	94	3.58	52
3	62.8	356	0.49	59	28	76.0	609	1.96	61
4	69.7	579	1.78	50	29	66.5	183	2.47	42
5	69.5	400	1.00	70	30	63.3	502	2.38	74
6	60.1	154	0.88	63	31	72.6	1,182	2.35	66
7	76.4	1,234	1.24	70	32	76.4	7,216	3.97	68
8	70.1	343	2.08	55	33	65.3	512	2.08	59
9	60.0	186	1.87	60	34	73.1	1,236	2.59	50
10	65.6	65	2.29	64	35	76.0	2,247	2.56	61
11	66.5	788	3.07	70	36	57.8	65	0.91	72
12	66.5	600	2.98	60	37	80.1	157	1.66	53
13	71.0	871	2.32	58	38	66.6	423	1.96	64
14	68.3	562	3.07	62	39	59.8	429	2.11	54
15	68.7	337	1.33	59	40	64.2	487	1.06	58
16	66.1	296	3.70	76	41	67.7	218	2.08	60
17	56.4	126	1.99	54	42	74.6	190	2.62	59
18	60.9	252	2.74	62	43	67.9	169	1.03	50
19	66.4	439	2.08	60	44	72.7	1,422	1.87	69
20	72.2	558	3.43	65	45	65.5	269	1.15	73
21	63.3	379	1.99	63	46	72.4	531	2.77	68
22	72.6	453	1.24	61	47	78.7	421	3.76	45
23	70.1	233	2.86	37	48	61.9	235	2.56	80
24	56.2	194	1.24	65	49	85.6	1,866	1.90	37
25	74.8	435	4.00	58	50	58.1	88	1.87	59

(continued)

Cont.	OverCost	NumEmp	Size	PerCosts	Cont.	OverCost	NumEmp	Size	PerCosts
51	75.6	1,833	4.45	55	69	60.4	127	0.76	63
52	63.0	870	1.54	66	70	80.9	3,766	3.19	55
53	67.9	946	2.29	56	71	74.8	1,576	3.52	54
54	65.8	422	4.72	65	72	79.2	764	3.04	57
55	57.1	79	2.74	64	73	68.1	408	1.36	50
56	74.7	393	4.54	64	74	66.8	370	1.57	70
57	66.1	229	1.66	68	75	83.6	769	2.23	53
58	68.5	316	3.07	57	76	61.7	1,041	3.01	63
59	55.2	224	1.54	66	77	76.2	546	2.86	63
60	60.9	573	1.09	70	78	64.3	147	1.27	51
61	72.3	461	2.50	66	79	71.3	148	1.72	55
62	70.2	732	1.48	68	80	63.8	501	1.42	57
63	62.2	189	2.02	64	81	80.4	1,686	2.26	57
64	58.1	195	2.29	65	82	80.1	1,264	2.68	58
65	66.2	962	2.17	60	83	59.9	229	0.43	67
66	84.1	964	4.90	45	84	65.5	111	0.28	57
67	81.6	921	3.28	54	85	73.0	2,138	3.82	63
68	76.7	214	2.62	79	86	67.0	356	3.58	55

a. Obtain correlations of all pairs of variables. Is there a severe collinearity problem with the data?
b. Plot overhead costs against each of the other variables. Locate a possible high influence outlier.
c. Obtain a regression equation (with overhead costs as the dependent variable) using all the data including any potential outlier.
d. Delete the potential outlier, and get a revised regression equation. How much did the slopes change?

Gov. **12.60** Consider the outlier-deleted regression model of Exercise 12.59.
a. Locate the F statistic. What null hypothesis is being tested? What can we conclude based on the F statistic?
b. Locate the t statistic for each independent variable. What conclusions can we reach based on the t tests?

Gov. **12.61** Use the outlier-deleted data of Exercise 12.59 to predict overhead costs of a contract when the contractor has 500 employees, the contract is 2.50% of the contractor's income, and personnel costs are 55% of direct costs. Obtain a 95% prediction interval. Would overhead costs equal to 88.9% of direct costs be unreasonable in this situation?

Bus. **12.62** The owner of a rapidly growing computer store tried to explain the increase in biweekly sales of computer software, using four explanatory variables: number of titles displayed, display footage, current customer base of Windows-based computers, and current customer base of Mac computers. The data are given below for the 52 biweekly sales periods as follows: biweek, software sales, number of titles displayed, display footage, Windows-based customers, and Mac based customers.

Biweek	Sales	Titles	Footage	Windows	Mac	Biweek	Sales	Titles	Footage	Windows	Mac
1	86.7	116	78	362	179	6	91.7	115	77	349	168
2	86.0	122	89	318	197	7	80.7	110	66	330	153
3	76.6	112	70	306	154	8	85.2	113	72	360	182
4	87.6	116	79	337	166	9	106.6	129	93	354	206
5	90.4	122	90	354	184	10	91.4	124	82	381	183

(continued)

Biweek	Sales	Titles	Footage	Windows	Mac	Biweek	Sales	Titles	Footage	Windows	Mac
11	105.0	125	85	387	201	32	120.8	148	104	441	251
12	102.3	131	96	387	203	33	125.9	161	104	500	270
13	94.0	125	85	346	201	34	128.4	164	111	501	277
14	93.6	122	78	339	173	35	127.9	163	109	471	274
15	109.6	140	101	418	211	36	126.7	164	110	483	282
16	108.2	136	92	409	211	37	120.2	163	107	494	261
17	107.9	139	99	414	210	38	115.1	162	105	456	253
18	108.4	138	96	440	226	39	115.9	167	117	474	275
19	89.2	127	75	382	188	40	134.1	172	125	512	275
20	92.6	134	88	383	211	41	131.8	174	117	520	296
21	104.4	129	80	397	213	42	140.1	170	111	507	275
22	107.6	134	88	384	227	43	157.1	175	119	517	264
23	107.2	141	91	407	242	44	152.7	178	124	499	297
24	102.6	141	89	434	217	45	136.5	173	115	474	278
25	104.7	138	84	424	211	46	140.4	170	109	515	296
26	112.2	145	98	428	245	47	130.8	171	112	482	274
27	115.9	145	99	415	237	48	121.7	167	104	510	282
28	113.3	147	103	443	251	49	124.7	173	115	488	294
29	109.1	142	91	414	217	50	138.6	179	127	539	294
30	112.8	145	98	412	250	51	148.4	188	134	578	325
31	111.1	142	92	424	217	52	142.0	183	123	536	302

a. Before doing the calculations, consider the economics of the situation, and state what sign you would expect for each of the partial slopes.
b. Obtain a multiple regression equation with sales as the dependent variable and all other variables as independent. Does each partial slope have the sign you expected in part (a)?
c. Calculate a 95% confidence interval for the coefficient of the titles variable. The computer output should contain the calculated standard error for this coefficient. Does the interval include 0 as a plausible value?

Gov. **12.63** **a.** In the regression model of Exercise 12.62, can the null hypothesis that none of the variables has predictive value be rejected at normal α levels?
b. According to t tests, which predictors, if any, add statistically detectable predictive value ($\alpha = .05$) given all the others?

Gov. **12.64** Obtain correlation coefficients for all pairs of variables from the data of Exercise 12.62. How severe is the collinearity problem in the data?

Gov. **12.65** Compare the coefficient of determination (R^2) for the regression model of Exercise 12.62 to the square of the correlation between sales and titles in Exercise 12.64. Compute the incremental F statistic for testing the null hypothesis that footage, Windows base, and Mac base add no predictive value given titles. Can this hypothesis be rejected at $\alpha = .01$?

Bus. **12.66** The market research manager of a catalog clothing supplier has begun an investigation of what factors determine the typical order size the supplier receives from customers. From the sales records stored on the company's computer, the manager obtained average order size data for 180 zip code areas. A part-time intern looked up the latest census information on per capita income, average years of formal education, and median price of an existing house in each of these zip code areas. (The intern couldn't find house price data for two zip codes and entered 0 for those areas.) The manager also was curious whether climate had any bearing on order size and included data on the average daily high temperature in winter and in summer.

The market research manager has asked for your help in analyzing the data. The output provided is intended only as a first try. The manager would like to know whether there was any evidence that the temperature variables mattered much and also which of the other variables seemed useful. There is some question about whether putting in 0 for the missing house price data was the right thing to do or whether that might distort the results. Please provide a basic, not-too-technical explanation of the results in this output and any other analyses you choose to perform.

```
MTB > name c1 'AvgOrder' c2 'Income' c3 'Educn' &
CONT> c4 'HousePr' c5 'WintTemp' c6 'SummTemp'
MTB > correlations of c1-c6

          AvgOrder   Income   Educn  HousePr WintTemp
Income       0.205
Educn        0.171    0.913
HousePr      0.269    0.616   0.561
WintTemp    -0.134   -0.098   0.014    0.066
SummTemp    -0.068   -0.115   0.005    0.018    0.481

MTB > regress c1 on 5 variables in c2-c6

The regression equation is
AvgOrder = 36.2 + 0.078 Income - 0.019 Educn
  + 0.0605 HousePr - 0.223 WintTemp + 0.006 SummTemp

Predictor       Coef      Stdev    t-ratio       p
Constant       36.18      12.37       2.92   0.004
Income        0.0780     0.4190       0.19   0.853
Educn        -0.0189     0.5180      -0.04   0.971
HousePr      0.06049    0.02161       2.80   0.006
WintTemp     -0.2231     0.1259      -1.77   0.078
SummTemp      0.0063     0.1646       0.04   0.969

Analysis of Variance

SOURCE       DF          SS         MS       F        p
Regression    5      417.63      83.53    3.71    0.003
Error       174     3920.31      22.53
Total       179     4337.94

SOURCE      DF    SEQ SS
Income       1    182.94
Educn        1      7.18
HousePr      1    142.63
WintTemp     1     84.84
SummTemp     1      0.03

Unusual Observations
Obs.    Income AvgOrder      Fit  Stdev.Fit  Residual   St.Resid
 25      17.1   23.570   36.555      0.632   -12.985     -2.76R
 78      11.9   24.990   34.950      0.793    -9.960     -2.13R
 83      13.4   36.750   29.136      2.610     7.614      1.92X
 87      14.3   45.970   35.918      0.463    10.052      2.13R
111      11.1   21.720   33.570      0.802   -11.850     -2.53R
113      10.4   43.500   33.469      0.817    10.031      2.15R
143      16.1   20.350   27.915      3.000    -7.565     -2.06RX
149      13.2   44.970   35.369      0.604     9.601      2.04R
169      13.5   44.650   34.361      0.660    10.289      2.19R
180      13.7   23.050   34.929      0.469   -11.879     -2.51R
R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

12.67 The following data were taken from the article ***"Toxaemic Signs During Pregnancy" [Applied Statistics (1983) 32:69–72].*** The data given here relate signs of toxemia, the presence or absence of hypertension and proteinuria, for 13,384 pregnant women classified by social class and smoking habit. The aim of the research was to determine if the amount of smoking and social class

of the women were associated with the incidence of signs of toxemia. The explanatory variables were social class (I, II, III, IV, V), an ordinal-level variable, and level of smoking (1—none; 2—1 to 19 cigarettes per day; 3—20 or more cigarettes per day).

		Signs of Toxemia				
Social Class	Smoking Level	None	Hypertension Only	Proteinuria Only	Both Hypertension and Proteinuria	Total
I	1	286	21	82	28	417
I	2	71	5	24	5	105
I	3	13	0	3	1	17
II	1	785	34	266	50	1,135
II	2	284	17	92	13	406
II	3	34	3	15	0	52
III	1	3,160	164	1,101	278	4,703
III	2	2,300	142	492	120	3,054
III	3	383	32	92	16	523
IV	1	656	52	213	63	984
IV	2	649	46	129	35	859
IV	3	163	12	40	7	222
V	1	245	23	78	20	366
V	2	321	34	74	22	451
V	3	65	4	14	7	90

a. Determine a model to relate the probability of hypertension in a pregnant woman to social class and smoking level.
b. Predict the probability of hypertension in a pregnant woman of social class III smoking 20 or more cigarettes per day.
c. Place a 95% confidence interval on the probability of hypertension in a pregnant woman of social class III smoking 20 or more cigarettes per day.

12.68 Refer to Exercise 12.67.
a. Determine a model to relate the probability of proteinuria in a pregnant woman to social class and smoking level.
b. Predict the probability of proteinuria in a pregnant woman of social class I smoking less than 20 cigarettes per day.
c. Place a 95% confidence interval on the probability of proteinuria in a pregnant woman of social class I smoking less than 20 cigarettes per day.

12.69 Refer to Exercise 12.67.
a. Determine a model to relate the probability of both hypertension and proteinuria in a pregnant woman to social class and smoking level.
b. Predict the probability of both hypertension and proteinuria in a pregnant woman of social class II smoking 1–19 cigarettes per day.
c. Place a 95% confidence interval on the probability of both hypertension and proteinuria in a pregnant woman of social class II smoking 1–19 cigarettes per day.

12.70 Refer to Exercise 12.67.
a. Determine a model to relate the probability of a pregnant woman having neither hypertension nor proteinuria to social class and smoking level.
b. Predict the probability of a nonsmoking pregnant woman of social class III having neither hypertension nor proteinuria.
c. Place a 95% confidence interval on the probability of a nonsmoking pregnant woman of social class III having neither hypertension and proteinuria.

Bio. **12.71** Refer to the fishery data in Example 12.17. The researcher wanted to determine if the modeling of the number of bass caught in the lake was altered by whether or not there is public access to the lake.

- **a.** Fit a first-order model relating catch to residency, size, and structure with separate intercepts and slopes for those lakes with access and those without access (access is used as an indicator variable).
- **b.** Display the fitted regression lines for lakes both with and without access.
- **c.** Test at the $\alpha = .05$ level whether there is a significant difference between the partial slopes for residency, size, and structure for the lakes with and without access.

Bio. **12.72** Refer Exercise 12.71.

- **a.** Estimate the mean catch for a lake having residency = 70, size = .8, and structure = 80 for lakes both with and without access.
- **b.** Place 95% confidence intervals on both of your estimates. Comment on the differences between the estimates for lakes with and without access.

Bio. **12.73** Refer to the fishery data in Example 12.17.

- **a.** Fit a second-order model relating catch to residency, size, and structure (with the squared terms but without the cross-product terms) with separate intercepts and slopes for those lakes with access and those without access (access is used as an indicator variable).
- **b.** Display the fitted regression lines for lakes both with and without access.
- **c.** Test whether the squared terms in residency, size, and structure provide a significant improvement to the fit of the model compared to the model with just the first-order terms.

Bio. **12.74** Refer Exercise 12.73.

- **a.** Estimate the mean catch for a lake having residency = 70, size = .8, and structure = 80 for lakes both with and without access using the second-order model.
- **b.** Place 95% confidence intervals on both of your estimates. Comment on the differences between the estimates for lakes with and without access.
- **c.** Compare the intervals on the estimates from the second-order model to those from the first-order model.

Bio. **12.75** Refer to Example 12.17. Why would it not be possible to fit a complete second-order model to this data—that is, a model including the three explanatory variables and their squares, cross-products, and terms, allowing separate partial slope coefficients for the two types of lakes?

CHAPTER 13

Further Regression Topics

13.1 Introduction and Abstract of Research Study

In Chapter 12, we presented the background information needed to use multiple regression. We discussed the general linear model and its use in multiple regression and introduced the normal equations, a set of simultaneous equations used in obtaining least-squares estimates for the βs of a multiple regression equation. Next, we presented standard errors associated with the $\hat{\beta}_j$ and their use in inferences about a single parameter β_j, a set of βs, $E(y)$, and a future value of y. We also considered special situations—comparing the slopes of several regression lines and the logistic regression problem. Finally, we condensed all of these inferential techniques using matrices.

This chapter is devoted to putting multiple regression into practice. How does one begin to develop an appropriate multiple regression for a given problem? Although there are no hard and fast rules, we can offer a few hints.

First, for each problem, you must decide on the dependent variable and candidate independent variables for the regression equation. This selection process will be discussed in Section 13.2. In Section 13.3, we consider how one selects the form of the multiple regression equation. The final step in the process of developing a multiple regression is to check for violation of the underlying assumptions. Tools for assessing the validity of the assumptions will be discussed in Section 13.4.

Following these steps *once* for a given problem will not ensure that you have an appropriate model. Rather, the regression equation seems to evolve as these steps are applied repeatedly, depending on the problem. For example, having considered candidate independent variables (step 1) and selected the form for a regression model involving some of these variables (step 2), we may find that certain assumptions have been violated (step 3). This will mean that we may have to return to either step 1 or step 2, but, hopefully, we have learned from

our previous deliberations and can modify the variables under consideration and/or the model(s) selected for consideration. Eventually, a regression model will emerge that meets the needs of the experimenter. Then the analysis techniques of Chapter 12 can be used to draw inferences about model parameters $E(y)$ and y.

Research Study: Construction Costs for Nuclear Power Plants

Advocates for nuclear power state that this source of electrical power provides net environmental benefits. Under the assumption that carbon dioxide emissions are associated with global warming, nuclear power plants would be an improvement over fossil fuel–based power plants. There is considerably less air pollution from nuclear power plants in comparison to coal or natural gas plants with respect to the production of sulfur oxides, nitrogen oxides, or other particulates. The waste from a nuclear plant differs from the waste from fossil fuel–based plants in that it is a solid-waste, spent fuel and some process chemicals, steam, and heated cooling water. The volume and mass of the waste from a nuclear power plant are much smaller than those of the waste from a fossil fuel–based plant. Some fossil fuel–based emissions can be limited or managed through pollution control equipment. However, these types of devices greatly increase the cost of building or managing the power plant. Similarly, nuclear plant operators and managers must spend money to control the radioactive wastes from their plants. An environmental component of any decision between building a nuclear or a fossil fuel plant is the cost of such controls and how they might change the costs of building and operating the power plant. Controversial decisions must also be made regarding what controls are appropriate. As public concerns increase about the level of pollution from coal-powered plants and the diminishing availability of other fossil fuels, the resistance to the construction of nuclear power plants has been reduced.

One of the major issues confronting power companies in seeking alternatives to fossil fuels is the need to forecast the costs of constructing nuclear power plants. The data, presented in Table 13.13 at the end of this chapter, are from the book **Applied Statistics *(Cox and Snell, 1981)*** and provide information on the construction costs of 32 light water reactor (LWR) nuclear power plants. The data set also contains information on the construction of the plants and specific characteristics of each power plant. The research goal is to determine which of the explanatory variables are most strongly related to the capital cost of the plant. If a reasonable model can be produced from these data, then the construction costs of new plants meeting specified characteristics can be predicted. Because of the resistance of the public and politicians to the construction of nuclear power plants, there is only a limited amount of data associated with new construction. The data set provided by Cox and Snell has only $n = 32$ plants along with 10 explanatory variables. The book *Introduction to Regression Modeling* (Abraham and Ledolter, 2006) provides a detailed analysis of this data set. At the end of this chapter, we will document some of the steps needed to build a model and then assess its usefulness in predicting the cost of constructing specific types of nuclear power plants.

13.2 Selecting the Variables (Step 1)

Perhaps the most critical decision in constructing a multiple regression model is the initial selection of independent variables. In later sections of this chapter, we consider many methods for refining a multiple regression analysis, but first we must make a decision about which independent (x) variables to consider for inclusion—and hence which data to gather. If we do not have useful data, we are unlikely to come up with a useful predictive model.

Although initially it may appear that an optimum strategy might be to construct a monstrous multiple regression model with very many variables, such models are difficult to interpret and are much more costly from a data-gathering and analysis time standpoint. How can a researcher make a reasonable selection of initial variables to include in a regression analysis?

selection of the independent variables

Knowledge of the problem area is critically important in the initial selection of data. First, identify the dependent variable to be studied. Individuals who have had experience with this variable by observing it, trying to predict it, and trying to explain changes in it often have remarkably good insight as to what factors (independent variables) affect it. As a consequence, the first step involves consulting those who have the most experience with the dependent variable of interest. For example, suppose that the problem is to forecast the next quarter's sales volume of an inexpensive brand of computer printer for each of 40 districts. The dependent variable y is then district sales volume. Certain independent variables, such as the advertising budget in each district and the number of sales outlets, are obvious candidates. A good district sales manager undoubtedly could suggest others.

collinearity

A major consideration in selecting predictor variables is the problem of **collinearity**—that is, severely correlated independent variables. A partial slope in multiple regression estimates the predictive effect of changing one independent variable while holding all others constant. However, when some or all of the predictors vary together, it can be almost impossible to separate out the predictive effects of each one. A common result when predictors are highly correlated is that the overall F test is highly significant, but none of the individual t tests comes close to significance. The significant F result indicates only that there is detectable predictive value somewhere among the independent variables; the nonsignificant t-values indicate that we cannot detect *additional* predictive value for any variable given all the others. The reason is that highly correlated predictors are surrogates for each other; any of them individually may be useful, but adding others will not be. When seriously collinear independent variables are all used in a multiple regression model, it can be virtually impossible to decide which predictors are in fact related to the dependent variable.

correlation matrix

There are several ways to assess the amount of collinearity in a set of independent variables. The simplest method is to look at a (Pearson) **correlation matrix,** which can be produced by almost all computer packages. The higher these correlations, the more severe the collinearity problem is. In most situations, any correlation over .9 or so definitely indicates a serious problem.

scatterplot matrix

Some computer packages can produce a **scatterplot matrix,** a set of scatterplots for each pair of variables. Collinearity appears in such a matrix as a close linear relation between two of the *independent* variables. For example, a sample of automotive writers rated a new compact car on 0- to 100-point scales for performance, comfort, appearance, and overall quality. The promotion manager doing the study wanted to know which variables best predicted the writers' rating of overall quality. A Minitab scatterplot matrix is shown in Figure 13.1. There are clear linear relations among the performance, comfort, and appearance ratings, indicating substantial collinearity. The following matrix of correlations confirms that fact:

```
MTB > correlations c1-c4

          Correlations (Pearson)
         overall  perform  comfort
perform    0.698
comfort    0.769    0.801
appear     0.630    0.479    0.693
```

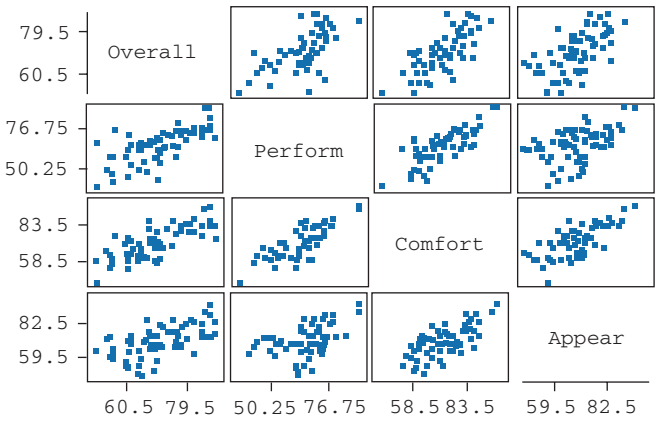


FIGURE 13.1
Scatterplot matrix for auto writers data

A scatterplot matrix can also be useful in detecting nonlinear relations or outliers. The matrix contains scatterplots of the dependent variable against each independent variable separately. Sometimes a curve or a serious outlier will be clear in the matrix. Other times the effect of other independent variables may conceal a problem. The analysis of residuals, discussed later in this chapter, is another good way to look for assumption violations.

The correlation matrix and scatterplot matrix may not reveal the full extent of a collinearity problem. Sometimes two predictors together predict a third all too well, even though either of the two by itself shows a more modest correlation with the third one. (Direct labor hours and indirect labor hours together predict total labor hours remarkably well, even if either one predicts the total imperfectly.) A number of more sophisticated ways of diagnosing collinearity are built into various computer packages. One such diagnostic is the variance inflation factor (VIF) discussed in Chapter 12.

A proposed full model uses k explanatory variables, $x_1, x_2, \ldots, x_k$, to explain the variation in the response variable, y:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

The VIF of the estimator of the jth partial slope, β_j, associated with the jth explanatory variable, x_j, is given by

$$\text{VIF}_j = 1/(1 - R_j^2)$$

where R_j^2 is the coefficient of determination from the regression of x_j on the remaining $k - 1$ explanatory variables. When x_j is linearly dependent on the other explanatory variables, the value of R_j^2 will be close to one, and VIF_j will be large. There is strong evidence of collinearity in the explanatory variables when the value of VIF exceeds 10. A detailed discussion of several diagnostic measures of collinearity can be found in the books by Cook and Weisberg (1982) and by Belsley, Kuh, and Welsch (1980).

EXAMPLE 13.1

Mercury contamination in freshwater fish has been a recognized problem in North America for over four decades. High concentrations of mercury in fish can pose a serious health threat to humans and birds. ***The paper "Influence of Water Chemistry on Mercury Concentration in Largemouth Bass from Florida Lake" (Lange, Royals, and Connor, 1993)*** evaluated the relationships between

mercury concentrations and selected physical and chemical lake characteristics. The researchers were attempting to determine if chemical characteristics of lakes strongly influenced the bioaccumulation of mercury in largemouth bass. The study included 53 lakes that were hydrologically diverse and spanned a wide range in terms of size and alkalinites. The data are given in Table 13.1.

TABLE 13.1
Mercury contamination data

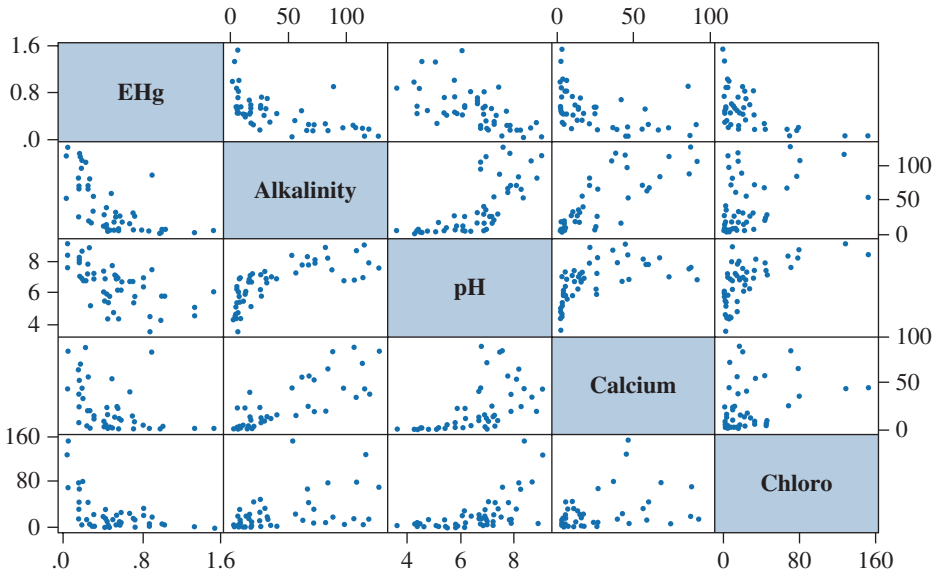
Lake	EHg	Alk	pH	Ca	Chlo	Lake	EHg	Alk	pH	Ca	Chlo
1	1.53	5.9	6.1	3	.7	28	.87	3.9	4.5	3.3	7
2	1.33	3.5	5.1	1.9	3.2	29	.50	5.5	4.8	1.7	14.8
3	.04	116	9.1	44.1	128.3	30	.47	6.3	5.8	3.3	.7
4	.44	39.4	6.9	16.4	3.5	31	.25	67	7.8	58.6	43.8
5	1.33	2.5	4.6	2.9	1.8	32	.41	28.8	7.4	10.2	32.7
6	.25	19.6	7.3	4.5	44.1	33	.87	5.8	3.6	1.6	3.2
7	.45	5.2	5.4	2.8	3.4	34	.56	4.5	4.4	1.1	3.2
8	.16	71.4	8.1	55.2	33.7	35	.16	119.1	7.9	38.4	16.1
9	.72	26.4	5.8	9.2	1.6	36	.16	25.4	7.1	8.8	45.2
10	.81	4.8	6.4	4.6	22.5	37	.23	106.5	6.8	90.7	16.5
11	.71	6.6	5.4	2.7	14.9	38	.04	53	8.4	45.6	152.4
12	.51	16.5	7.2	13.8	4	39	.56	8.5	7	2.5	12.8
13	.54	25.4	7.2	25.2	11.6	40	.89	87.6	7.5	85.5	20.1
14	1.00	7.1	5.8	5.2	5.8	41	.18	114	7	72.6	6.4
15	.05	128	7.6	86.5	71.1	42	.19	97.5	6.8	45.5	6.2
16	.15	83.7	8.2	66.5	78.6	43	.44	11.8	5.9	24.2	1.6
17	.19	108.5	8.7	35.6	80.1	44	.16	66.5	8.3	26	68.2
18	.49	61.3	7.8	57.4	13.9	45	.67	16	6.7	41.2	24.1
19	1.02	6.4	5.8	4	4.6	46	.55	5	6.2	23.6	9.6
20	.70	31	6.7	15	17	47	.27	81.5	8.9	20.5	9.6
21	.45	7.5	4.4	2	9.6	48	.98	1.2	4.3	2.1	6.4
22	.59	17.3	6.7	10.7	9.5	49	.31	34	7	13.1	4.6
23	.41	12.6	6.1	3.7	21	50	.43	15.5	6.9	5.2	16.5
24	.81	7	6.9	6.3	32.1	51	.58	25.6	6.2	12.6	27.7
25	.42	10.5	5.5	6.3	1.6	52	.28	17.3	5.2	3	2.6
26	.53	30	6.9	13.9	21.5	53	.25	71.8	7.9	20.5	8.8
27	.31	55.4	7.3	15.9	24.7						

The variables in Table 13.1 are as follows:

Lake	ID number of the lake
EHg	expected mercury concentration (mg/g) for a 3-year-old fish (inferred from data)
Alk	alkalinity level in lake (mg/L as $CaCO_3$)
pH	degree of acidity ($0 \le pH \le 7$) or alkalinity ($7 < pH \le 14$)
Ca	calcium level (mg/L)
Chlo	chlorophyll (mg/g)

A scatterplot matrix is shown in Figure 13.2 along with the pairwise correlation from Minitab. Is there any indication of collinearity in the four explanatory variables? Does the matrix plot suggest any other problems with the data?

FIGURE 13.2
Matrix plot of EHg, alkalinity, pH, calcium, chlorophyll



Correlations: Alkalinity, pH, Calcium, Chloro

	Alkalinity	pH	Calcium
pH	0.719		
Calcium	0.833	0.577	
Chloro	0.478	0.608	0.410

Solution The plots in Figure 13.2 indicate a positive linear relationship between alkalinity and pH and between alkalinity and calcium, with a somewhat weaker positive relationship between calcium and pH and between chlorophyll and pH. The relationships between chlorophyll and calcium and between alkalinity and chlorophyll are very weak. These observations are confirmed by the values from the correlation matrix. Based on the correlation values, the only pair of explanatory variables that would be of concern for collinearity would be calcium and alkalinity. However, with a correlation of 0.833, there is no indication of a serious collinearity problem in the data. Further, there appear to be two lakes that have data values that may be of high leverage. Lakes 3 and 38 have chlorophyll values that are considerably larger than the values for the remaining 51 lakes. As we discussed in Chapter 11, a data point that has high leverage may greatly influence the slope of the line relating mercury content to amount of chlorophyll in the lake. Also, the data value associated with lake 40 may have high influence in that in the plots of EHg versus alkalinity and EHg versus calcium, the EHg value for lake 40 is much larger than the EHg values for the other data points that have values for alkalinity and calcium similar to those for lake 40. ■

One of the best ways to avoid collinearity problems is to choose predictor variables intelligently, right at the beginning of a regression study. Try to find independent variables that should correlate decently with the dependent variable but do not have obvious correlations with each other. If possible, try to find independent variables that reflect various components of the dependent variable.

For example, suppose we want to predict the sales of inexpensive printers for personal computers in each of 40 sales districts. Total sales are made up of several sectors of buyers. We might identify the important sectors as college students, home users, small businesses, and computer network workstations. Therefore, we might try number of college freshmen, household income, small business starts, and new network installations as independent variables. Each one makes sense as a predictor of printer sales, and there is no obvious correlation among the predictors. People who are knowledgeable about the variable you want to predict can often identify components and suggest reasonable predictors for the different components.

EXAMPLE 13.2

A firm that sells and services desktop computers is concerned about the volume of service calls. The firm maintains several district service branches within each sales region, and computer owners requiring service call the nearest branch. The branches are staffed by technicians trained at the main office. The key problem is whether technicians should be assigned to main office duty or to service branches; assignment decisions have to be made monthly. The required number of service branch technicians grows in almost exact proportion to the number of service calls. Discussion with the service manager indicates that the key variables in determining the volume of service calls seem to be the number of computers in use, the number of new installations, whether or not a model change has been introduced recently, and the average temperature. (High temperatures, or possibly the associated high humidity, lead to more frequent computer troubles, especially in imperfectly air conditioned offices.) Which of these variables can be expected to correlate with the others?

Solution It is hard to imagine why temperature should be correlated with any of the other variables. There should be some correlation between number of computers in use and number of new installations, if only because every new installation is a computer in use. Unless the firm has been growing at an increasing rate, we would not expect a severe correlation (we would, however, like to see the data). The correlation of model change to number in use and new installations is not at all obvious; surely data should be collected and correlations analyzed. ■

A researcher who begins a regression study may try to put too many independent variables into a regression model; hence, we need some sensible guidelines to help select the independent variables to be included in the final regression model from potential candidates.

To sort out which independent variables should be included in a regression model from the list of variables generated from discussions with experts, you can resort to any one of a number of selection procedures. We will consider several of these in this text; for further details, consult Neter, Kutner, Nachtsheim, and Wasserman (1996).

The first selection procedure involves performing *all possible regressions* with the dependent variable and one or more of the independent variables from the list of candidate variables. Obviously, this approach should not be attempted unless the analyst has access to a computer with suitable software and sufficient core to run a large number of regression models relatively efficiently.

TABLE 13.2
Data on 20 independent pharmacies

PHARMACY	VOLUME	FLOOR—SP	PRESC—RX	PARKING	SHOPCNTR	INCOME
1	22	4,900	9	40	1	18
2	19	5,800	10	50	1	20
3	24	5,000	11	55	1	17
4	28	4,400	12	30	0	19
5	18	3,850	13	42	0	10
6	21	5,300	15	20	1	22
7	29	4,100	20	25	0	8
8	15	4,700	22	60	1	15
9	12	5,600	24	45	1	16
10	14	4,900	27	82	1	14
11	18	3,700	28	56	0	12
12	19	3,800	31	38	0	8
13	15	2,400	36	35	0	6
14	22	1,800	37	28	0	4
15	13	3,100	40	43	0	6
16	16	2,300	41	20	0	5
17	8	4,400	42	46	1	7
18	6	3,300	42	15	0	4
19	7	2,900	45	30	1	9
20	17	2,400	46	16	0	3

As an illustration, we will use hypothetical data on prescription sales data (volume per month) obtained for a random sample of 20 independent pharmacies. These data, along with data on the total floor space, percentage of floor space allocated to the prescription department, number of parking spaces available for the store, whether the pharmacy is in a shopping center, and per capita income for the surrounding community are recorded in Table 13.2.

Before running all possible regressions for the data of Table 13.2, we need to consider what criterion should be used to select the best-fitting equation from all possible regressions. The first and perhaps simplest criterion for selecting the best regression equation from the set of all possible regression equations involves computing an estimate of the error variance, σ_ε^2, using $s_\varepsilon^2 = \text{MS(Residual)} = \text{SS(Residual)}/[n - (k + 1)]$. Since this quantity is used in most inferences (statistical tests and confidence intervals) about model parameters and $E(y)$, it would seem reasonable to choose the model that has the smallest value of s_ε^2, the mean square error.

A second criterion makes use of the *coefficient of determination,* R^2, which is computed for each model. We then choose from amongst those models having highest R^2 values. There is a limitation in using this criterion. Suppose we denote the coefficient of determination computed from a model having k explanatory variables and an intercept term (that is, $k + 1$ regression coefficients) by R_k^2, where

$$R_k^2 = \frac{\text{SS(Total)} - \text{SS}_k\text{(Residual)}}{\text{SS(Total)}} = 1 - \frac{\text{SS}_k\text{(Residual)}}{\text{SS(Total)}}$$

where $SS_k(\text{Residual})$ is the residual sum of squares from a model with k explanatory variables and $SS(\text{Total}) = \sum_{i=1}^{n}(y_i - \bar{y})^2$. The term SS(Total) is the same for all models, but $SS_k(\text{Residual})$ may be quite different depending on k, and, furthermore, even for the same k there may be many different models having the same number of explanatory variables but in different combinations. Consider the five explanatory variables in Table 13.2. There are 10 different models in which the model contains three of the five explanatory variables. We would thus have 10 different values of R_3^2 in this case. In selecting amongst the 10 models using three of the five explanatory variables, we generally would prefer the model having the largest value for R_3^2. In general, if we increase the number of explanatory variables in the model, then SS(Residual) decreases or stays the same. By increasing the number of explanatory variables in the model, we can eventually obtain a model in which R_k^2 is very close to one. In fact, if we have n data values and the model contains n regression coefficients, then SS(Residual) = 0 and $R_n^2 = 1$. Thus, R^2 can lead to misleading results if we are trying to balance the two criteria of obtaining a model in which we have a good fit and obtaining one in which we have a limited number of explanatory variables.

adjusted R^2

For the reasons given above, we will define an **adjusted R^2,** which provides for a penalty for each regression coefficient included in the model:

$$R^2_{adj,\,k} = 1 - \frac{SS_k(\text{Residual})/(n-k-1)}{SS(\text{Total})/(n-1)} = 1 - \frac{(n-1)}{(n-k-1)}(1 - R_k^2)$$

Note that in $R^2_{adj,\,k}$, the sums of squares are adjusted for their corresponding degrees of freedom. Also, increasing the number of terms in the model from k to $k + 1$ will not always result in an increase in $R^2_{adj,\,k}$, as would be true for R_k^2. If the additional term does not result in a decrease in SS(Residual), then $R^2_{adj,\,k}$ will actually decrease, whereas R_{k+1}^2 would always be larger or the same as R_k^2. Thus, we will be penalized with a smaller $R^2_{adj,\,k}$ for including variables in the model that do not provide a reasonable improvement to the fit of the model to the data.

With one more algebraic manipulation, we can show that

$$R^2_{adj,\,k} = 1 - \frac{SS_k(\text{Residual})/(n-k-1)}{SS(\text{Total})/(n-1)} = 1 - \frac{s_\varepsilon^2}{SS(\text{Total})/(n-1)} = 1 - \frac{s_\varepsilon^2}{s_y^2}$$

From these two forms for R^2_{adj}, we can observe that the adjusted coefficient of determination is comparing the variability in the response variable without any explanatory variables, s_y^2, to the variability that remains in the ys after fitting a model to ys that includes k explanatory variables. Thus, selecting models using the criterion of a large value of R^2_{adj} is equivalent to selecting models using the criterion of a small value for s_ε^2.

EXAMPLE 13.3

Refer to the data of Table 13.2. Use the R^2_{adj} criterion to determine the best-fitting regression equation for one, two, three, and four independent variables.

Solution SAS output is provided here, and the regression equations with the highest R^2_{adj} values are summarized in Table 13.3.

```
SAS OUTPUT
Dependent Variable: VOLUME
Variable Selection
Number of Observations Read 20

Number    Adjusted
in Model  R-Square  R-Square   C(p)    AIC    BIC    Variables in Model

   3      0.6327    0.6907    2.4364  57.03  61.82  FLOOR_SP PRESC_RX SHOPCNTR
   2      0.6263    0.6657    1.6062  56.59  60.12  FLOOR_SP PRESC_RX
   3      0.6193    0.6794    2.9635  57.75  62.21  FLOOR_SP PRESC_RX PARKING
   4      0.6184    0.6987    4.0623  58.51  64.37  FLOOR_SP PRESC_RX PARKING SHOPCNTR
   4      0.6115    0.6933    4.3177  58.87  64.52  FLOOR_SP PRESC_RX SHOPCNTR INCOME
   2      0.6055    0.6471    2.4744  57.67  60.86  PRESC_RX SHOPCNTR
   3      0.6039    0.6664    3.5713  58.54  62.66  FLOOR_SP PRESC_RX INCOME
   3      0.5993    0.6626    3.7496  58.77  62.79  PRESC_RX PARKING SHOPCNTR
   4      0.5954    0.6806    4.9097  59.67  64.86  FLOOR_SP PRESC_RX PARKING INCOME
   5      0.5930    0.7001    6.0000  60.42  67.19  FLOOR_SP PRESC_RX PARKING SHOPCNTR INCOME
   3      0.5809    0.6471    4.4720  59.67  63.29  PRESC_RX SHOPCNTR INCOME
   4      0.5731    0.6630    5.7301  60.75  65.31  PRESC_RX PARKING SHOPCNTR INCOME
   3      0.5279    0.6024    6.5577  62.05  64.66  PRESC_RX PARKING INCOME
   2      0.4943    0.5475    7.1224  62.64  64.32  PRESC_RX INCOME
   2      0.4763    0.5314    7.8722  63.34  64.81  PRESC_RX PARKING
   2      0.4364    0.4958    9.5366  64.81  65.86  SHOPCNTR INCOME
   1      0.4082    0.4393   10.1709  64.93  65.87  PRESC_RX
   3      0.4064    0.5001   11.3332  66.63  67.45  FLOOR_SP SHOPCNTR INCOME
   3      0.4042    0.4983   11.4193  66.71  67.50  PARKING SHOPCNTR INCOME
   4      0.3683    0.5013   13.2789  68.59  69.14  FLOOR_SP PARKING SHOPCNTR INCOME
   2      0.1691    0.2565   20.7035  72.57  71.67  FLOOR_SP SHOPCNTR
   2      0.1449    0.2349   21.7147  73.15  72.12  FLOOR_SP INCOME
   3      0.1273    0.2651   22.3051  74.34  72.66  FLOOR_SP PARKING SHOPCNTR
   3      0.1161    0.2557   22.7427  74.60  72.84  FLOOR_SP PARKING INCOME
   2      0.1120    0.2054   23.0890  73.90  72.71  PARKING INCOME
   1      0.1007    0.1480   23.7702  73.30  72.82  INCOME
   1      -.0122    0.0411   28.7618  75.66  74.84  SHOPCNTR
   1      -.0202    0.0335   29.1129  75.82  74.97  FLOOR_SP
   2      -.0410    0.0686   29.4780  77.08  75.26  FLOOR_SP PARKING
   1      -.0505    0.0048   30.4539  76.41  75.48  PARKING
   2      -.0706    0.0421   30.7126  77.64  75.71  PARKING SHOPCNTR
```

TABLE 13.3
Best-fitting models, based on R^2_{adj}

Number of Explanatory Variables in Model	R^2_{adj}	Variables
1	0.4082	Prescription sales
2	0.6263	Floor space, prescription sales
3	0.6327	Shopping center, floor space, prescription sales
4	0.6184	Parking, shopping center, floor space, prescription sales
5	0.5930	Parking, shopping center, floor space, prescription sales, income

Although there is a sizable increase in R^2_{adj} when the number of explanatory variables is increased from one to two, there is very little improvement by including three or four explanatory variables. Therefore, the best overall model based on R^2_{adj}, considering both number of variables and fit of the model, would be the model containing the variables floor space and prescription sales.

The SAS output displays the values for R^2. This example illustrates the problem in using R^2 as a measure of the best model. Examining the values for R^2, it can be seen that the models with the highest R^2 values are the models with four variables, then with three variables, and so on. However, when using the values of R^2_{adj}, the three models with largest R^2_{adj} values are two three-variable models and a two-variable model, not one of the five four-variable models.

Keep in mind that the object of our search is to choose the subset of independent variables that generates the best prediction equation for *future* values of y; unfortunately, however, because we do not know these future values, we focus on criteria that choose the best-fitting regression equations to the known sample y-values. One possible bridge between this emphasis on the best fit to the known sample y-values and that on choosing the best predictor of future y-values is to split the sample data into two parts—one part used for fitting the various regression equations and the other part used for validating how well the prediction equations can predict ''future'' values. Although there is no universally accepted rule for deciding how many of the data should be included in the ''fitting'' portion of the sample and how many go into the ''validating'' portion of the sample, it is reasonable to split the total sample in half provided the total sample size n is greater than $2p + 20$, where p is the number of parameters in the largest potential regression model. A possible criterion for the best prediction equation would involve minimizing $\sum(y_i - \hat{y}_i)^2$ for the validating portion of the total sample.

Once the regression model is selected from the data-splitting approach, the entire set of sample data is used to obtain the final prediction equation. Thus, even though it appears we would only use part of the data, the entire data set is used to obtain the final prediction equation.

Observations do cost money, however, and it may be impractical to obtain enough observations to apply the data-splitting approach for choosing the best-fitting regression equation. In these situations, a form of validation can be accomplished using the PRESS statistic. For a sample of y-values and a proposed regression model relating y to a set of xs, we first remove the first observation and fit the model using the remaining $n - 1$ observations. Based on the fitted equation, we estimate the first observation (denoted by $\hat{y}^*_1$) and compute the residual $y_1 - \hat{y}^*_1$. This process is repeated $n - 1$ times, successively removing the second, third, . . . , nth observation, each time computing the residual for the removed observation. The PRESS statistic is defined as

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \hat{y}^*_i)^2$$

The model that gives the smallest value for the PRESS statistic is chosen as the best-fitting model.

To this point, we have considered criteria for selecting the best-fitting regression model from a subset of independent variables. In general, if we choose a model that leaves out one or more ''important'' predictor variables, our model is *underspecified*, and the additional variability in the y-values that would be accounted for with these variables becomes part of the estimated error variance. At the other end of the spectrum, if we choose a model that contains one or more ''extraneous'' predictor variables, our model is *overspecified*, and we stand the chance of having a *multicollinearity* problem. We will deal with this problem later. The point is that a criterion, based on the C_p statistic, seems to balance some pros and cons of previously presented selection criteria, along with the problems of over- and

underspecification, to arrive at a choice of the best-fitting subset regression equation. The C_p statistic (see Mallows, 1973) is

$$C_p = \frac{\text{SS(Residual)}_p}{s_\varepsilon^2} - (n - 2p)$$

where SS(Residual)_p is the sum of squares for error from a model with p parameters (including β_0) and s_ε^2 is the mean square error from the regression equation with the largest number of independent variables. For a given selection problem, compute C_p for every regression equation that is fit. Theory suggests that the best-fitting model should have $C_p \approx p$. For a model with k explanatory variables, $p = k + 1$.

EXAMPLE 13.4

Refer to the output of Example 13.3. Determine the value of C_p for all possible regressions with one, two, three, and four independent variables. Select the best-fitting equation for one, two, three, and four independent variables. Which regression equation seems to give the best overall fit, based on the C_p statistic?

Solution The best-fitting models are summarized in Table 13.4. Based on the C_p criterion, there would be very little difference between the best-fitting models for three and four independent variables. The most "important" predictive variables would be parking space and prescription sales because they appear in the best-fitting models for three and four independent variables. Note that the important independent variables found in Example 13.3 are different from the ones related by C_p.

TABLE 13.4 Best-fitting models, C_p criterion

Number of Independent Variables	p	C_p	Variables
1	2	10.17	Prescription sales
2	3	1.61	Floor space, prescription sales
		2.47	Prescription sales, shopping center
3	4	3.75	Prescription sales, parking space, shopping center
4	5	4.91	Floor space, prescription sales, parking, income
5	6	6.00	All five independent variables

Two other criteria for selecting the most crucial independent variables are based on information criteria. The Akaike's information criterion (AIC) selects the model having the smallest value of AIC:

$$\text{AIC}_k = n\log_e(\text{SS(Residual)}/n) + 2k$$

AIC balances the model selection process between the goodness of fit of the model, as measured by SS(Residual), and the model complexity, the number of terms in the model, $2k$. As the number of terms in the model increases, SS(Residual) decreases, but the penalty of model complexity, $2k$, increases.

Thus, if a large number of unnecessary independent variables were placed in the model that resulted in only a small reduction in SS(Residual), the model complexity penalty, the value of $2k$, would exceed this decrease in SS(Residual), resulting in an increase in AIC, not a decrease. Chatterjee and Hadi (2012) recommend that models with AIC not differing by more than two units should be treated as equally adequate.

Sheather (2009) states that when either the sample size is small or the number of parameters in the model divided by the sample size, k/n, is relatively large, using AIC to select the independent variables in the model will tend to put too many terms in the model, referred to as overfitting the model.

An alternative to AIC was proposed by Schwarz (1978). The Bayesian Information Criterion (BIC) selects the model having the smallest value of BIC:

$$\text{BIC}_k = n\log_e(\text{SS(Residual)}/n) + k\log_e(n)$$

AIC and BIC differ with respect to the model complexity penalty, $2k$ versus $k\log_e(n)$, with BIC placing a much more severe penalty for overfitting the model. For smaller data sets, BIC will often correct the tendency of AIC to overfit the model; that is, BIC will generally place fewer terms in the model than will AIC.

EXAMPLE 13.5

Refer to the SAS output of Example 13.3. Determine the value of AIC and BIC for all possible regression models with one, two, three, and four independent variables. Select the best-fitting model for one, two, three, and four independent variables. Which model seems to produce the best overall fit, based on AIC? Answer this question using BIC also.

Solution The best-fitting models are summarized below for each value of k, the number of independent variables in the model. Based on AIC and using the recommendation that models not differing by two units should be treated as equally adequate, the models with two, three, and four variables as listed below would be considered as the best-fitting models. Based on BIC, the four-variable model would have a somewhat higher BIC value than the two- and three-variable models. This would confirm the observation that AIC tends to overfit models when the sample size is small. An overall recommendation based on combining the values of C_p, AIC, and BIC would be a three-variable model, However, the variables selected using C_p differ from the variables selected by AIC and BIC. The variables prescription sales and parking are in common, but C_p would include shopping center, whereas AIC and BIC would include floor space.

k	AIC	BIC	Independent Variables in Model
1	64.93	65.87	Prescription sales
2	56.59	60.12	Floor space, prescription sales
3	57.03	61.82	Floor space, prescription sales, shopping center
4	58.51	64.37	Floor space, prescription sales, parking, shopping center
5	60.42	67.19	Floor space, prescription sales, parking, shopping center, income

Best subset regression provides another procedure for finding the best-fitting regression equation from a set of k candidate independent variables. This procedure uses an algorithm that avoids running all possible regressions. The computer program prints a listing of the best M (the user selects M) regression equations with one independent variable in the model, two independent variables in the model, three independent variables in the model, and so on, up to the model containing all k independent variables in the model. Some programs allow the user to specify the criterion for "best" (for example, C_p or maximum R^2_{adj}), whereas other programs fix the criterion.

EXAMPLE 13.6

Use the SAS output in Example 13.3 to find the $M = 2$ best subset regression equations of size one to five based on the AIC criterion for the data of Table 13.2. From the various "best" regression equations, select the regression equation that has the "best" AIC.

Solution The relevant information is given below. There are two best subsets of size one to four. Based on the maximum R^2, the subset with all independent variables will always be the best regression. However, based on AIC, BIC, adjusted R^2, or C_p, our conclusion would differ from the best obtained from the maximum R^2.

```
SAS OUTPUT
Dependent Variable: VOLUME
Variable Selection
Number of Observations Read 20

Number    Adjusted
in Model  R-Square  R-Square    C(p)     AIC     BIC    Variables in Model

   1      0.4082    0.4393   10.1709   64.93   65.87   PRESC_RX
   1      0.1007    0.1480   23.7702   73.30   72.82   INCOME
   2      0.6263    0.6657    1.6062   56.59   60.12   FLOOR_SP PRESC_RX
   2      0.6055    0.6471    2.4744   57.67   60.86   PRESC_RX SHOPCNTR
   3      0.6327    0.6907    2.4364   57.03   61.82   FLOOR_SP PRESC_RX SHOPCNTR
   3      0.6193    0.6794    2.9635   57.75   62.21   FLOOR_SP PRESC_RX PARKING
   4      0.6184    0.6987    4.0623   58.51   64.37   FLOOR_SP PRESC_RX PARKING SHOPCNTR
   4      0.6115    0.6933    4.3177   58.87   64.52   FLOOR_SP PRESC_RX SHOPCNTR INCOME
   5      0.5930    0.7001    6.0000   60.42   67.19   FLOOR_SP PRESC_RX PARKING SHOPCNTR INCOME
```

backward elimination
stepwise regression

A number of other procedures can be used to select the best regression, and although we will not spend a great deal more time on this subject, we will mention briefly the **backward elimination** method and **stepwise regression** procedure.

The backward elimination method begins with fitting the regression model that contains all the candidate independent variables. For each independent variable x_j, we compute

$$F_j = \frac{\text{SSR}_j - \text{SSR}}{\text{MS(Residual)}} \qquad j = 1, 2, \ldots$$

where SSR is the sum-of-squares residuals from the complete model and SSR_j is the sum-of-squares residuals from the model that contains all xs except x_j. MS(Residual) is the mean square error for the complete model. Let min F_j denote the smallest F_j value. If min $F_j < F_{\alpha}$, where α is the preselected significance level, remove the independent variable corresponding to min F_j from the regression equation. The backward elimination process then begins all over again with one variable removed from the list of candidate independent variables. Thus, backward elimination starts with the complete model with all independent variables entered and eliminates variables one at a time until a reasonable candidate regression model is found. This occurs when, in a particular step, min $F_j > F_{\alpha}$; the resulting complete model is the best-fitting regression equation.

Stepwise regression, on the other direction, works in the other direction, starting with the model $y = \beta_0 + \varepsilon$ and adding variables one at a time until a stopping criterion is satisfied. At the initial stage of the process, the first variable entered into the equation is the one with the largest F test for regression. At the second stage, the two variables to be included in the model are the variables with the largest F test for regression of two variables. Note that the variable entered in the first step might not be included in the second step; that is, the best single variable might not be one of the best two variables. Because of this, some people use a simplified stepwise regression (sometimes called *forward selection*) whereby, once a variable is entered, it cannot be eliminated from the regression equation at a later stage.

EXAMPLE 13.7

Use the data of Example 13.3 to find the variables to be included in a regression equation based on backward elimination. Comment on your findings.

Solution SAS output is shown for a backward elimination procedure applied to the data of Table 13.2. As indicated, backward elimination begins with all (five) candidate variables in the regression equation. This is designated as step 0 in the backward elimination process. Then one by one, independent variables are eliminated until min $F_j > F_{\alpha}$. Note that in step 1, the variable income is removed and in step 2, the variable parking is removed from the regression equation. Step 3 is the final step in the process for this example; the variable shopping center is removed. As indicated in the output, the remaining variables comprise the best-fitting regression equation based on backward elimination. That equation is

$$\hat{y} = 48.291 - .004(\text{floor space}) - .582(\text{prescription sales})$$

which is identical to the result we obtained from the other variable selection procedures.

```
REGRESSION ANALYSIS, USING BACKWARD ELIMINATION

Backward Elimination Procedure for Dependent Variable VOLUME

Step 0    All Variables Entered     R-square = 0.70007369   C(p) =  6.00000000
```

```
                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       5          525.44030541     105.08806108       6.54   0.0025
Error           14          225.10969459      16.07926390
Total           19          750.55000000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP      42.08710826     10.43775070     261.42703544      16.26   0.0012
FLOOR_SP      -0.00241878      0.00183889      27.81923726       1.73   0.2095
PRESC_RX      -0.50046955      0.16429694     149.19783807       9.28   0.0087
PARKING       -0.03690284      0.06546687       5.10907792       0.32   0.5819
SHOPCNTR      -3.09957355      3.24983522      14.62673442       0.91   0.3564
INCOME         0.10666360      0.42742012       1.00135642       0.06   0.8066

Bounds on condition number:     7.823107,     117.1991
--------------------------------------------------------------------------------

Step 1   Variable INCOME Removed    R-square = 0.69873952   C(p) = 4.06227626

                DF         Sum of Squares       Mean Square          F   Prob>F

Regression       4          524.43894899      131.10973725       8.70   0.0008
Error           15          226.11105101       15.07407007
Total           19          750.55000000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP      43.46782063      8.56960161      387.83321233      25.73   0.0001
FLOOR_SP      -0.00228513      0.00170330       27.13112543       1.80   0.1997
PRESC_RX      -0.52910174      0.11386382      325.48983690      21.59   0.0003
PARKING       -0.03952477      0.06256589        6.01580808       0.40   0.5371
SHOPCNTR      -2.71387948      2.76799605       14.49041122       0.96   0.3424

Bounds on condition number:     5.071729,      46.98862
--------------------------------------------------------------------------------

Step 2   Variable PARKING Removed   R-square =  0.69072432  C(p) = 2.43641080

                DF         Sum of Squares       Mean Square          F   Prob>F

Regression       3          518.42314091      172.80771364      11.91   0.0002
Error           16          232.12685909       14.50792869
Total           19          750.55000000

                Parameter        Standard          Type II
Variable         Estimate           Error    Sum of Squares          F   Prob>F

INTERCEP      42.82702645      8.34803435      381.83242065      26.32   0.0001
FLOOR_SP      -0.00247284      0.00164539       32.76871130       2.26   0.1523
PRESC_RX      -0.52941361      0.11170410      325.87978038      22.46   0.0002
SHOPCNTR      -3.03834296      2.66836223       18.81002755       1.30   0.2716

Bounds on condition number:     4.917388,      30.31995
--------------------------------------------------------------------------------

Step 3   Variable SHOPCNTR Removed  R-square = 0.66566267    C(p) = 1.60624219

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       2          499.61311336     249.80655668      16.92   0.0001
Error           17          250.93688664      14.76099333
Total           19          750.55000000
```

```
                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP      48.29085530      6.89043477     725.02357305      49.12   0.0001
FLOOR_SP      -0.00384228      0.00113262     169.87259933      11.51   0.0035
PRESC_RX      -0.58189034      0.10263739     474.44587802      32.14   0.0001

Bounds on condition number:    2.290122,      9.160487
------------------------------------------------------------------------------

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination Procedure for Dependent Variable VOLUME

        Variable   Number    Partial     Model
Step    Removed        In      R**2      R**2       C(p)           F   Prob>F

   1    INCOME          4    0.0013    0.6987     4.0623      0.0623   0.8066
   2    PARKING         3    0.0080    0.6907     2.4364      0.3991   0.5371
   3    SHOPCNTR        2    0.0251    0.6657     1.6062      1.2965   0.2716
```

EXAMPLE 13.8

Describe the results of stepwise regression applied to the data of Table 13.2.

Solution The SAS output for the data of Table 13.2 is shown here. Stepwise regression begins with the model $y = \beta_0 + \varepsilon$ and adds variables one at a time. For these data, the variable prescription sales was entered in step 1 of the stepwise procedure, the variable floor space was added to the regression model in step 2, and the variable shopping center was added in step 3. No other variables met the entrance criterion of $p = .5$ for inclusion in the model. If the criterion was more selective, requiring a relatively small p-value (say, .15 or less) for each new independent variable, the stepwise regression procedure would not include the variable shopping center in step 3 (with a p-value of .2716), and we would arrive at the same best-fitting regression equation that we obtained previously with other methods.

```
REGRESSION ANALYSIS, USING FORWARD ELIMINATION

Forward Selection Procedure for Dependent Variable VOLUME

Step 1   Variable PRESC_RX Entered  R-square = 0.43933184   C(p) = 10.17094219

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       1          329.74051403     329.74051403      14.10   0.0014
Error           18          420.80948597      23.37830478
Total           19          750.55000000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP      25.98133346      2.58814791    2355.90463660     100.77   0.0001
PRESC_RX      -0.32055657      0.08535423     329.74051403      14.10   0.0014

Bounds on condition number:           1,           1
------------------------------------------------------------------------------
```

```
Step 2   Variable FLOOR_SP Entered  R-square = 0.66566267    C(p) = 1.60624219

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       2           499.61311336     249.80655668      16.92   0.0001
Error           17           250.93688664      14.76099333
Total           19           750.55000000

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares          F   Prob>F

INTERCEP      48.29085530      6.89043477     725.02357305      49.12   0.0001
FLOOR_SP      -0.00384228      0.00113262     169.87259933      11.51   0.0035
PRESC_RX      -0.58189034      0.10263739     474.44587802      32.14   0.0001

Bounds on condition number:     2.290122,         9.160487
--------------------------------------------------------------------------------

Step 3  Variable SHOPCNTR Entered  R-square = 0.69072432     C(p) = 2.43641080

                DF         Sum of Squares      Mean Square          F   Prob>F

Regression       3           518.42314091     172.80771364      11.91   0.0002
Error           16           232.12685909      14.50792869
Total           19           750.55000000

                Parameter        Standard          Type II
Variable         Estimate           Error     Sum of Squares          F   Prob>F

INTERCEP      42.82702645      8.34803435       381.83242065     26.32   0.0001
FLOOR_SP      -0.00247284      0.00164539        32.76871130      2.26   0.1523
PRESC_RX      -0.52941361      0.11170410       325.87978038     22.46   0.0002
SHOPCNTR      -3.03834296      2.66836223        18.81002755      1.30   0.2716

Bounds on condition number:    4.917388,        30.31995
--------------------------------------------------------------------------------

No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection Procedure for Dependent Variable VOLUME

        Variable   Number   Partial     Model
Step    Entered        In      R**2      R**2        C(p)          F  Prob>F

   1    PRESC_RX        1    0.4393    0.4393     10.1709     14.1046  0.0014
   2    FLOOR_SP        2    0.2263    0.6657      1.6062     11.5082  0.0035
   3    SHOPCNTR        3    0.0251    0.6907      2.4364      1.2965  0.2716
```

In a typical regression problem, you ascertain which variables are potential candidates for inclusion in a regression model (step 1) by discussing the problem with experts and/or by using any one of a number of possible selection procedures. For example, we could run all possible regressions, apply a best subset regression approach, or follow a stepwise regression (or backward elimination) procedure. This list is by no means exhaustive. Sometimes the various criteria do single out the same model as best (or near best, as seen with the data of Table 13.2). At other times, you may get different models from the different criteria. Which approach is best? Which one should we believe and use?

The most important response to these questions is that with the availability and accessibility of a computer and applicable software systems, it is possible

to work effectively with any of these selection procedures; no one procedure is universally accepted as better than the others. Hence, rather than attempting to use some or all of the procedures, you should begin to use one method (perhaps because of the availability of particular software in your computer facility) and learn as much as you can about it by continued use. Then you will be well equipped to solve almost any regression problem to which you are exposed.

13.3 Formulating the Model (Step 2)

In Section 13.2, we suggested several ways to develop a list of candidate independent variables for a given regression problem. We can and should seek the advice of experts in the subject matter area to provide a starting point, and we can employ any one of several selection procedures to come up with a possible regression model. In this section, we refine the information gleaned from step 1 to develop a useful multiple regression model.

Having chosen a subset of k independent variables to be candidates for inclusion in the multiple regression and the dependent variable y, we still may not know the actual relationship between the dependent and independent variables. Suppose the assumed regression model is of a lower order than is the actual model relating y to $x_1, x_2, \ldots, x_k$. Then provided there is more than one observation per factor–level combination of the independent variables, we can conduct a test of the inadequacy of a fitted polynomial model using the equation $F = \text{MS}_{\text{Lack}}/\text{MSP}_{\text{exp}}$ as discussed in Chapter 11.

Another way to examine an assumed (fitted) model for lack of fit is to examine scatterplots of residuals $(y_i - \hat{y}_i)$ versus x_j. For example, suppose that step 1 has indicated that the variables x_1, x_2, and x_3 constitute a reasonable subset of independent variables to be related to a response y using a multiple regression equation. Not knowing which polynomial function of the independent variables to use, we could start by fitting the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

to obtain the least-squares prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$. A plot of the residuals $(y_i - \hat{y}_i)$ versus each one of the xs would shed some light as to which higher-degree terms may be appropriate. We'll illustrate the concepts using residuals by way of a regression problem for one independent variable and then extend the concepts to a multiple regression situation.

EXAMPLE 13.9

In a radioimmunoassay, a hormone with a radioactive trace is added to a test tube containing an antibody that is specific to that hormone. The two will combine to form an antigen–antibody complex. To measure the extent of the reaction of the hormone with the antibody, we measure the amount of hormone that is bound to the antibody relative to the amount remaining free. Typically, experimenters measure the ratio of the bound/free radioactive count (y) for each dose of hormone (x) added to a test tube. Frequently, the relation between y and x is nearly linear. Data from 11 test tubes in a radioimmunoassay experiment are shown in Table 13.5.

TABLE 13.5
Radioimmunoassay data

Bound/Free Count	Dose (concentration)
9.900	.00
10.465	.25
10.312	.50
13.633	.75
20.784	1.00
36.164	1.25
62.045	1.50
78.327	1.75
90.307	2.00
97.348	2.25
102.686	2.50

a. Plot the sample data and fit the linear regression model

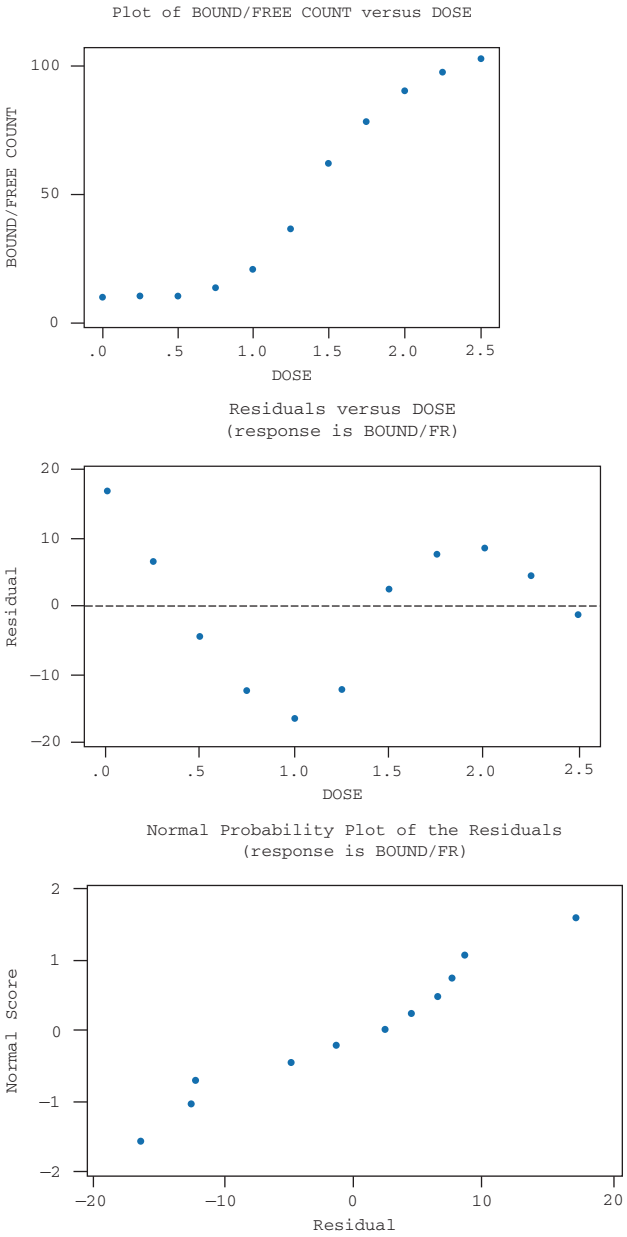
$$y = \beta_0 + \beta_1 x + \varepsilon$$

b. Plot the residuals versus count and versus $\hat{y}$. Does a linear model adequately fit the data?

c. Suggest an alternative (if appropriate).

Solution Computer output is shown here.

```
Data Display

Row  BOUND/FREE  COUNT  DOSE  DOSE_2
  1              9.900  0.00  0.0000
  2             10.465  0.25  0.0625
  3             10.312  0.50  0.2500
  4             13.633  0.75  0.5625


Row  BOUND/FREE  COUNT  DOSE  DOSE_2
  5             20.784  1.00  1.0000
  6             36.164  1.25  1.5625
  7             62.045  1.50  2.2500
  8             78.327  1.75  3.0625
  9             90.307  2.00  4.0000
 10             97.348  2.25  5.0625
 11            102.686  2.50  6.2500

Regression Analysis: BOUND/FREE COUNT versus DOSE

The regression equation is
BOUND/FREE COUNT = -7.19 + 44.4 DOSE

Predictor        Coef    SE Coef          T        P
Constant       -7.189      6.226      -1.15    0.278
DOSE           44.440      4.210      10.56    0.000

S = 11.04       R-Sq = 92.5%    R-Sq(adj) = 91.7%

Analysis of Variance

Source            DF         SS          MS        F        P
Regression         1      13577       13577   111.44    0.000
Residual Error     9       1097         122
Total             10      14674
```



a, b. The linear fit is

$$\hat{y} = -7.189 + 44.440x$$

The plot of y (count) versus x (concentration) clearly shows a lack of fit of the linear regression model; the residual plots confirm this same lack of fit. The linear regression underestimates counts at the lower and upper ends of the concentration scale and overestimates at the middle concentrations.

c. A possible alternative model would be a quadratic model in concentration:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

More will be said about this later in the chapter. ■

Scatterplots are not very helpful in detecting interactions among the independent variables other than for the two-independent variable case. The reason is that there are too many variables for most practical problems and it is difficult to present the interrelationships among independent variables and their joint effects on the response y using two-dimensional scatterplots. Perhaps the most reasonable suggestion is to use one of the best subset regression methods of the previous section, some trial-and-error fitting of models using the candidate independent variables, and a bit of common sense to determine which interaction terms should be used in the multiple regression model.

The presence of dummy variables (for qualitative independent variables) presents no major problem for ascertaining the adequacy of the fit of a polynomial model. The important thing to remember is that when quantitative and dummy variables are included in the same regression model, for each setting of the dummy variables, we obtain a regression in the quantitative variables. Hence, plotting methods for detecting an inadequate fit should be applied separately for each setting of the dummy variables. By examining these plots carefully, we can also detect potential differences in the forms of the polynomial models for different settings of the dummy variables.

EXAMPLE 13.10

A nutritional study involved participants taking a course in which they were given information concerning how to control their caloric intake. The study was conducted with 29 subjects aged 20 to 53 years, all of whom were healthy but moderately overweight. The researchers collected data on caloric intake during a 4-week period prior to the participants attending the course. During a second 4-week period 6 months after completing the course, the researchers once again collected information on caloric intake. The data in Table 13.6 provide information on the gender and age of the participants, along with the mean daily caloric intake prior to instruction and the percentage reduction in mean caloric intake during the second 4-week test period.

TABLE 13.6
Caloric intake data

Subject	Gender	Age	Before	Reduction
1	F	20	1,160	8.23
2	F	22	1,888	7.56
3	F	24	1,861	7.23
4	F	27	1,649	6.89
5	F	28	2,463	5.47
6	F	31	1,934	3.78
7	F	35	2,211	2.43
8	F	37	2,320	2.51
9	F	38	2,352	3.12
10	F	39	2,693	3.26
11	F	40	2,236	4.30
12	F	41	2,072	4.54
13	F	46	2,026	5.28
14	F	47	1,991	5.92
15	F	52	1,552	6.92
16	F	53	1,406	7.83
17	M	22	3,678	5.93

Subject	Gender	Age	Before	Reduction
18	M	23	3,101	5.10
19	M	26	3,418	8.19
20	M	32	2,891	2.00
21	M	33	2,273	4.75
22	M	33	2,509	2.71
23	M	34	3,689	3.64
24	M	36	2,789	3.65
25	M	37	3,018	2.75
26	M	42	2,754	2.84
27	M	45	2,567	4.23
28	M	47	2,177	2.43
29	M	49	2,695	2.18

The researchers were interested in studying the relationship between the percentage reduction in caloric intake and the explanatory variables: gender, age, and caloric intake prior to instruction. Fit a linear regression model and use residual plots to determine what (if any) higher-order terms are required. Do the same conclusions hold for males and females? Make suggestions for additional terms in the multiple regression model.

Solution A linear model in the three explanatory variables was fit to the data:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon$$

where

y = percentage reduction in caloric intake

$$x_1 = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

x_2 = age of participant

x_3 = caloric intake before instruction

From the SAS output, the estimated regression equation is

$$\hat{y} = 6.41 + 7.51x_1 - .115x_2 + .000531x_3 + .091x_1x_2 - .00441x_1x_3$$

Substituting $x_1 = 0$ and 1 into this equation, we obtain the separate regression equations for males and females, respectively:

$x_1 = 0$ (males)

$$\hat{y} = 6.41 - .115x_2 + .000531x_3$$

$x_1 = 1$ (females)

$$\hat{y} = 13.92 - .024x_2 - .00388x_3$$

Scatterplots of y versus x_2 and x_3 show that reduction in caloric intake decreases as male participants' ages increase but show a quadratic relationship for female participants. For female participants, reduction in caloric intake tends to decrease as the before caloric intake increases with the opposite relationship holding true for males.

LINEAR REGRESSION OF ENERGY INTAKE ON BEFORE AND AGE

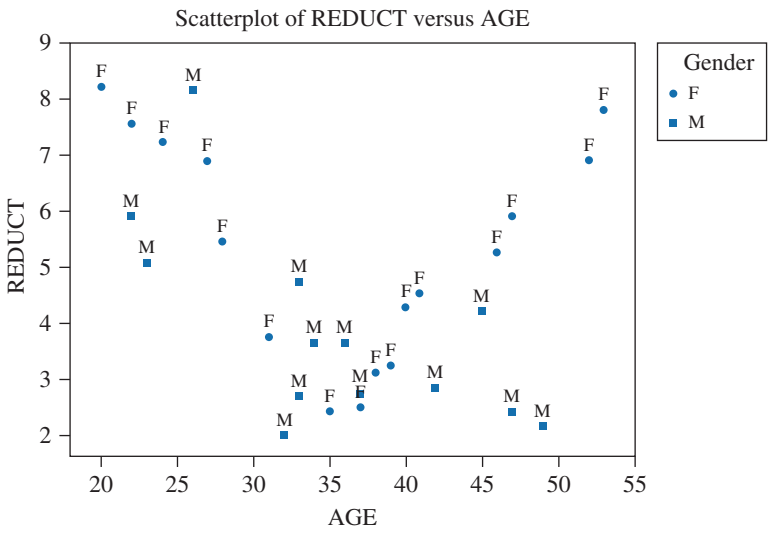
Dependent Variable: PERCENT CALORIC REDUCTION

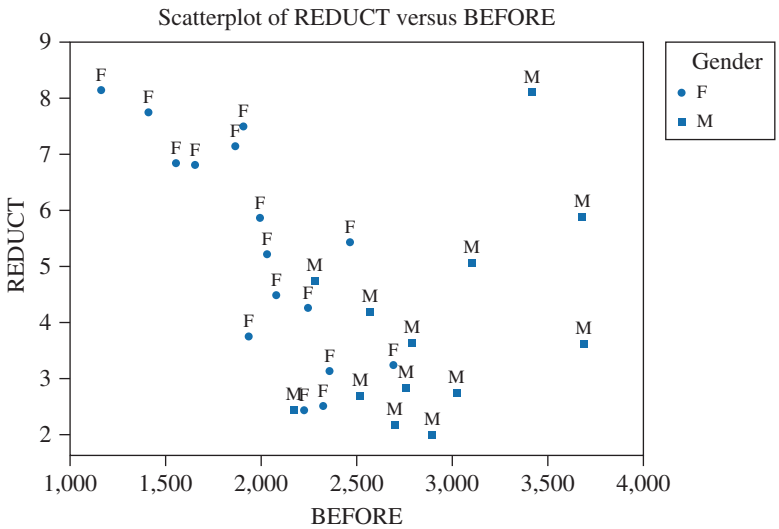
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	70.15325	14.03065	7.93	0.0002
Error	23	40.68257	1.76881		
Corrected Total	28	110.83581			

Root MSE	1.32997	R-Square	0.6329
Dependent Mean	4.67828	Adj R-Sq	0.5532
Coeff Var	28.42853		

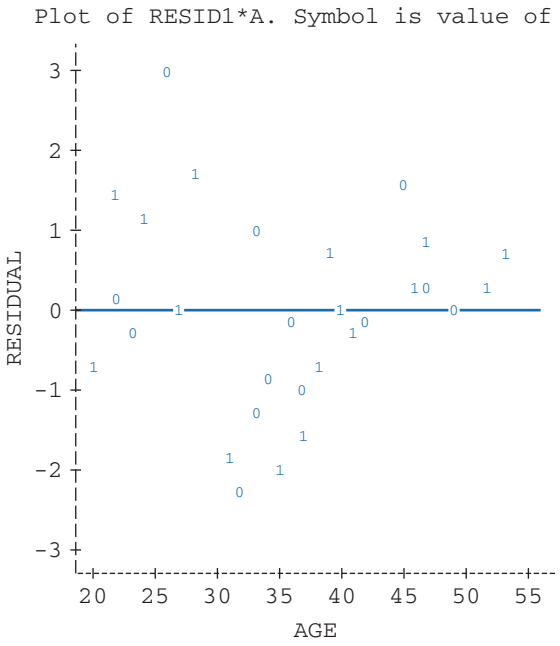
Parameter Estimates

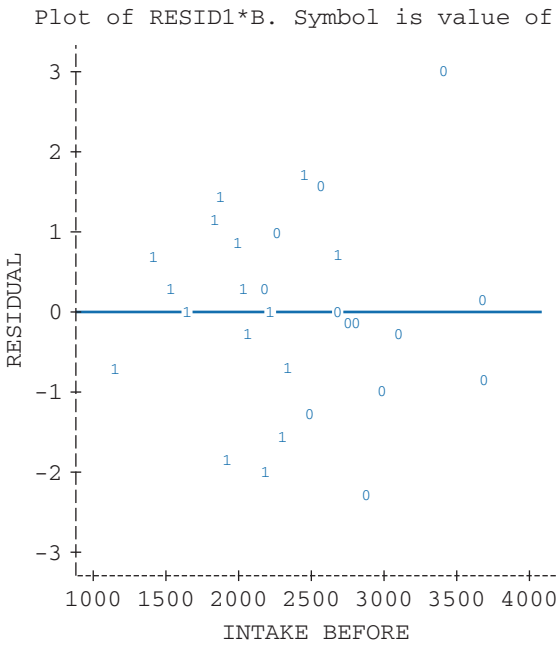
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	6.40924	4.50491	1.42	0.1682
I	GENDER	1	7.51016	4.95060	1.52	0.1429
A	AGE	1	-0.11521	0.05683	-2.03	0.0544
AI	GENDER*AGE	1	0.09091	0.06594	1.38	0.1812
B	INTAKE BEFORE	1	0.00053148	0.00102	0.52	0.6076
BI	GENDER*INTAKE BEFORE	1	-0.00441	0.00133	-3.32	0.0030





Residual plots from linear model





The residual plots versus age show an underestimation for middle-aged males and females but an overestimation for younger and older males and females. The residual plots versus caloric intake before did not reveal any discernable

patterns for either males or females. A second-order model in both x_2 and x_3 was fit to the data. Based on the plots, the quadratic terms in x_3 were probably unnecessary.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_3 + \beta_5 x_3^2 + \beta_6 x_1 x_2 + \beta_7 x_1 x_2^2 + \beta_8 x_1 x_3 + \beta_9 x_1 x_3^2 + \varepsilon$$

From the SAS output, the estimated regression equation is

$$\hat{y} = 22.664 + 1.604x_1 - .517x_2 + .00559x_2^2 - .00581x_3 + .00000104x_3^2 - .834x_1x_2 + .0125x_1x_2^2 + .0108x_1x_3 - .00000235x_1x_3^2 + \varepsilon$$

Substituting $x_1 = 0$ and 1 into this equation, we obtain the separate regression equations for males and females, respectively:

$x_1 = 0$ (males)

$$\hat{y} = 22.664 - .517x_2 + .00559x_2^2 - .00581x_3 + .00000104x_3^2$$

$x_1 = 1$ (females)

$$\hat{y} = 24.268 - 1.351x_2 + .0181x_2^2 + .00499x_3 - .00000131x_3^2$$

From the output from the two models, note that R^2_{adj} has increased from .5532 for the linear model to .6701 for the quadratic model. There has been a sizable increase in the fit of the model to the data.

QUADRATIC REGRESSION OF ENERGY INTAKE ON BEFORE AND AGE

Dependent Variable: PERCENT CALORIC REDUCTION

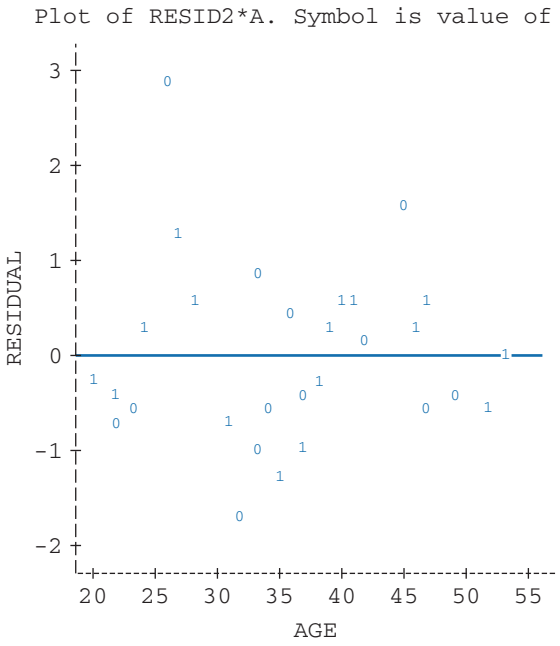
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	86.02731	9.55859	7.32	0.0001
Error	19	24.80850	1.30571		
Corrected Total	28	110.83581			

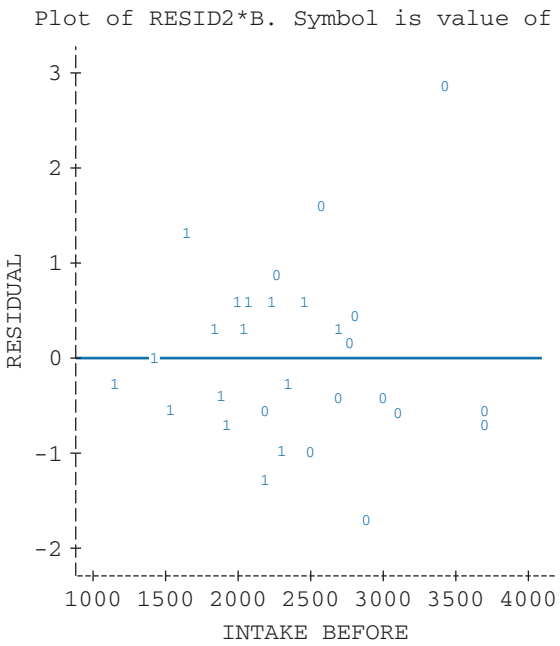
Root MSE	1.14268	R-Square	0.7762
Dependent Mean	4.67828	Adj R-Sq	0.6701
Coeff Var	24.42517		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	22.66395	13.57467	1.67	0.1114
I	INDICATOR FOR GENDER	1	1.60406	14.97921	0.11	0.9158
A	AGE	1	-0.51678	0.33877	-1.53	0.1436
A2	AGE SQUARED	1	0.00559	0.00465	1.20	0.2441
AI	AGE*GENDER	1	-0.83449	0.54323	-1.54	0.1410
A2I	AGE SQUARED*GENDER	1	0.01249	0.00741	1.69	0.1082
B	INTAKE BEFORE	1	-0.00581	0.00868	-0.67	0.5110
BI	INTAKE BEFORE*GENDER	1	0.01078	0.01105	0.98	0.3417
B2	INTAKE BEFORE SQUARED	1	0.00000104	0.00000146	0.71	0.4848
B2I	INTAKE SQUARED*GENDER	1	-0.00000235	0.00000219	-1.07	0.2980

Residual plots from quadratic model





So far in this section, we have considered lack of fit only as it relates to polynomial terms and interaction terms. However, sometimes the lack of fit is related not to the fact that we have not included enough higher-degree terms and interactions in the model but rather to the fact that y is not adequately represented by any polynomial model in the subset of independent variables.

FIGURE 13.3
Plots depicting nonlinear relationships

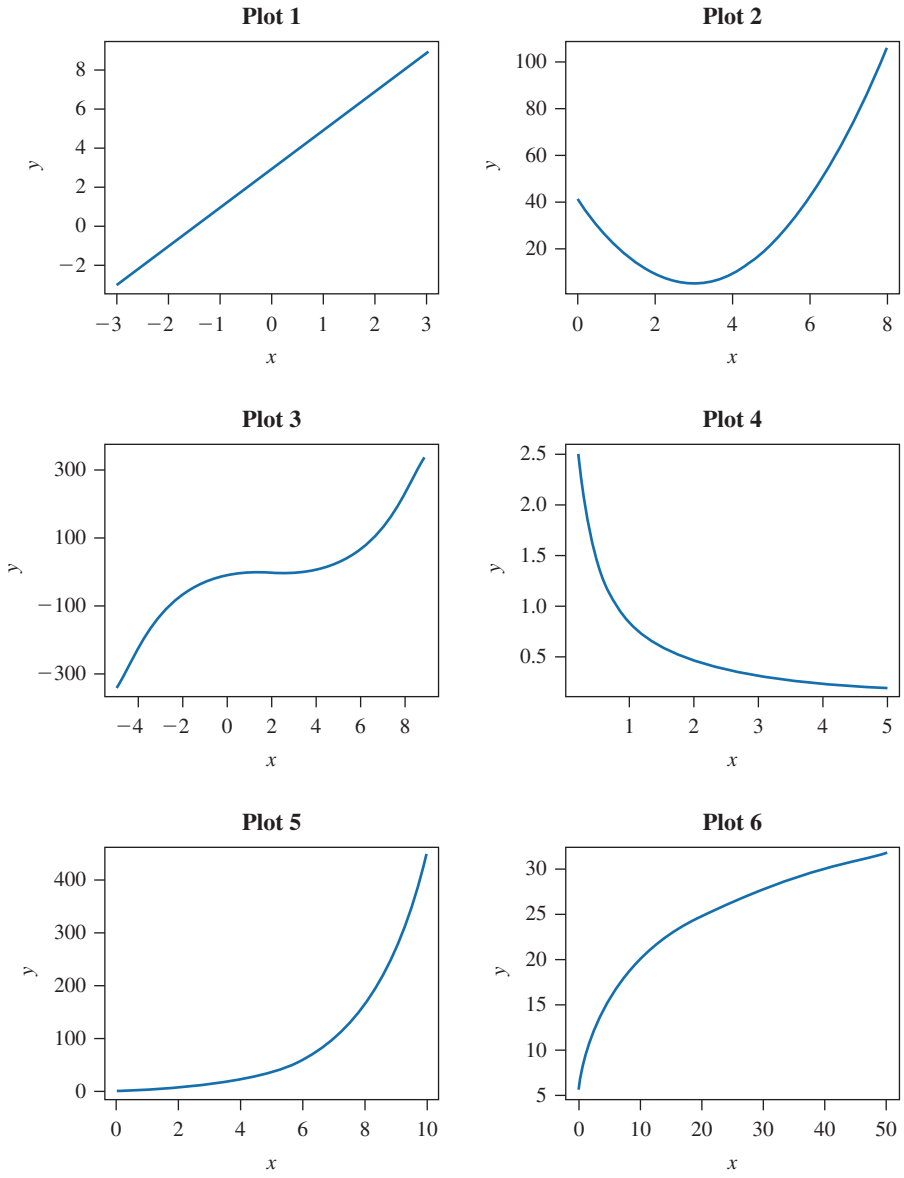


Figure 13.3 contains six plots of various functions of a response variable y to a single explanatory variable x:

The plots were generated using the following relationships between y and x:

Plot 1: $y = 2x + 3$

Plot 2: $y = 4(x - 3)^2 + 5$

Plot 3: $y = (x - 2)^3 + .6$

Plot 4: $y = \dfrac{1}{x + .2}$

Plot 5: $y = 3e^{x/2}$

Plot 6: $y = 8\log(x + 2)$

Plots 1–3 of Figure 13.3 demonstrate the great flexibility in the shape of models using a polynomial relationship between y and x. However, polynomial relationships do not cover all possible relationships unless we are willing to use a very high-order model, such as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_k x^k + e$$

where k is a very large integer. Plots 4–6 display shapes that can be obtained by using models involving negative exponents, exponentiation, or the log function. There may be situations in which a model that is *nonlinear* in the βs may be appropriate. Such models are displayed in plots 4–6 with the following general forms given here:

$$\text{Plot 4:} \quad y = \frac{1}{\beta_1 x + \beta_2}$$

$$\text{Plot 5:} \quad y = \beta_1 e^{x/\beta_2}$$

$$\text{Plot 6:} \quad y = \beta_1 \log(\beta_2 x + \beta_3)$$

In engineering problems, nonlinear models often arise as the solution of differential equations that govern an engineering process. In biological studies, nonlinear models often are used for growth models. Some examples of the application of nonlinear models in economics and finance will be presented next.

Most basic finance books show that if a quantity y grows at a rate r per unit time (continuously compounded), the value of y at time t is

$$y_t = y_0 e^{rt}$$

logarithmic transformation

where y_0 is the initial value. This relation may be converted into a linear relation between y_t and t by a **logarithmic transformation:**

$$\log y_t = \log y_0 + rt$$

The simple linear regression methods of Chapter 11 can be used to fit data for this regression model with $\beta_0 = \log y_0$ and $\beta_1 = r$. When y is an economic variable such as total sales, the logarithmic transformation is often used in a multiple regression model:

$$\log y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

The Cobb–Douglas production function is another standard example of a nonlinear model that can be transformed into a regression equation:

$$y = c l^{\alpha} k^{\beta}$$

where y is production, l is labor input, k is capital input, and α and β are unknown constants. Again, to transform the dependent variable, we take logarithms to obtain

$$\begin{aligned} \log y &= (\log c) + \alpha(\log l) + \beta(\log k) \\ &= \beta_0 + \beta_1(\log l) + \beta_2(\log k) \end{aligned}$$

which suggests that a regression of log production on log labor and log capital is linear.

EXAMPLE 13.11

In studying the relationships between the streams of people migrating to urban areas and the size of the urban areas, demographers have used a gravity-type model:

$$M = \frac{\alpha_1 S_1 S_2}{D^{\alpha_2}}$$

where α_1 and α_2 are unknown constants, M is the level of migration (interaction) between two urban areas, D is the distance from one urban area to the second urban area, and S_1 and S_2 are the population sizes of the two urban areas. Express this model as a linear model.

Solution By taking the natural logarithm of both sides of the equation, we would have

$$\log(M) = \log(\alpha_1) + \log(S_1) + \log(S_2) - \alpha_2 \log(D)$$

This model then can be expressed in a general form as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where $y = \log(M), \beta_0 = \log(\alpha_1), x_1 = \log(S_1), x_2 = \log(S_2), x_3 = \log(D)$, and $\beta_3 = -\alpha_2$. Data on M, S_1, S_2, and D would be needed in order to obtain estimates of the two constants, α_1 and α_2. ■

A logarithmic transformation is only one possibility. It is, however, particularly useful because logarithms convert a multiplicative relation to an additive one.

Another transformation that is sometimes useful is an inverse transformation, $1/y$. If, for instance, y is speed in meters per second, then $1/y$ is time required in seconds per meter. This transformation works well with very severe curvature; a logarithm works well with moderate curvature. Try them both; it is easy with a computer package. Another transformation that is particularly useful when a dependent variable increases to a maximum and then decreases, is a quadratic x^2 term. In this transformation, do not replace x by x^2; use them both as predictors. The same use of both x and x^2 works well if a dependent variable decreases to a minimum and then increases. A fairly extensive discussion of possible transformations is found in Tukey (1977).

The remaining material in this section should be considered optional. We will use computer software and output to illustrate the fitting of nonlinear models. The logic behind what we are doing is the same used in the least-squares method for the general linear model; in fact, the procedure is sometimes called **nonlinear least squares.** The sum of squares for error is defined as before,

nonlinear least squares

$$\text{SS(Residual)} = \sum_i (y_i - \hat{y}_i)^2$$

The problem is to find a method for obtaining estimates $\hat{\alpha}_1, \hat{\alpha}_2, \ldots$ that will minimize SS(Residual). The set of simultaneous equations used for finding these estimates is again called the set of normal equations, but unlike least squares for the general linear model, the form of the normal equations depends on the form of the nonlinear model being used. Also, because the normal equations involve nonlinear functions of the parameters, their solutions can be quite complicated. Because of this technical difficulty, a number of iterative methods have been developed for obtaining a solution to the normal equations.

For those of you with a background in calculus, the normal equations for a nonlinear model involve partial derivatives of the nonlinear function with respect to each of the parameters α_i. Fortunately, most of the computer software packages currently marketed (for example, SAS, SPSS, R, and JMP) approximate the derivative and do not require one to give the form of the normal equations; only the form of the nonlinear equation is needed. We will illustrate this with the data from a previous example.

EXAMPLE 13.12

In Example 13.9, we fit the model $y = \beta_0 + \beta_1 x + \varepsilon$ to the radioimmunoassay data. The residual plots from this fit suggested that higher-order terms in x were needed in the model. Fit a quadratic model, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$, to the data, and assess the fit.

Solution SAS output from fitting the model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ is shown here. From the residual plot, there appears to be a cyclical pattern in the residuals. This would indicate that the quadratic model did not provide an adequate fit and hence that an alternative model may be needed. When there is a cyclical pattern in the data, polynomial models do not generally provide an adequate fit.

A nonlinear model that may provide a more reasonable fit to the data is the following model:

$$y = \frac{\beta_0 - \beta_3}{1 + (x/\beta_2)^{\beta_1}} + \beta_3$$

where the parameters have the following interpretations:

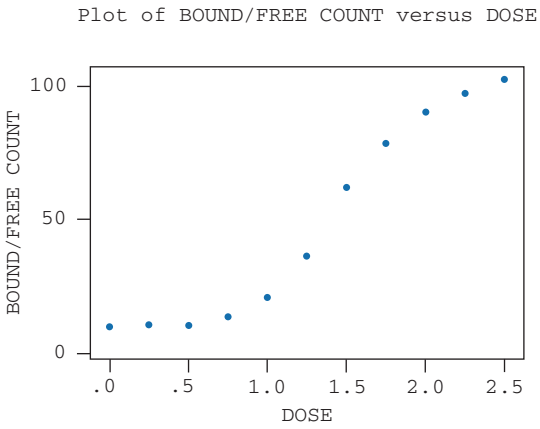
β_0: value of y at the lower end of the curve

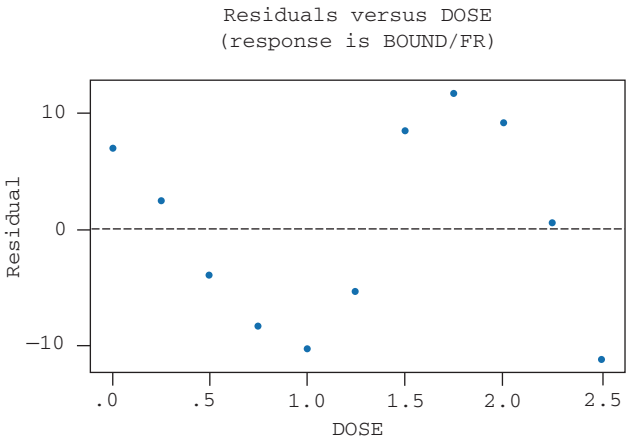
β_3: value of y at the upper end of the curve

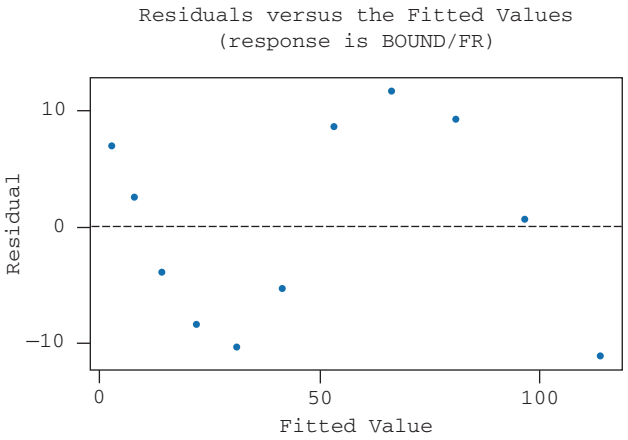
β_1: value of x corresponding to the value of y midway between β_0 and β_3

β_2: a slope-type measure

```
Regression Analysis: BOUND/FREE COUNT versus DOSE, DOSE_2

The regression equation is
BOUND/FREE COUNT = 2.88 + 17.6 DOSE + 10.7 DOSE_2

Predictor        Coef     SE Coef          T        P
Constant        2.884       7.175       0.40    0.698
DOSE            17.58       13.35       1.32    0.225
DOSE_2         10.745       5.144       2.09    0.070

S = 9.418       R-Sq = 95.2%     R-Sq(adj) = 94.0%

Analysis of Variance

Source            DF          SS          MS         F      P
Regression         2     13964.4      6982.2     78.72  0.000
Residual Error     8       709.6        88.7
Total             10     14674.0

Source  DF        Seq SS
DOSE     1       13577.4
DOSE_2   1         386.9
```







■

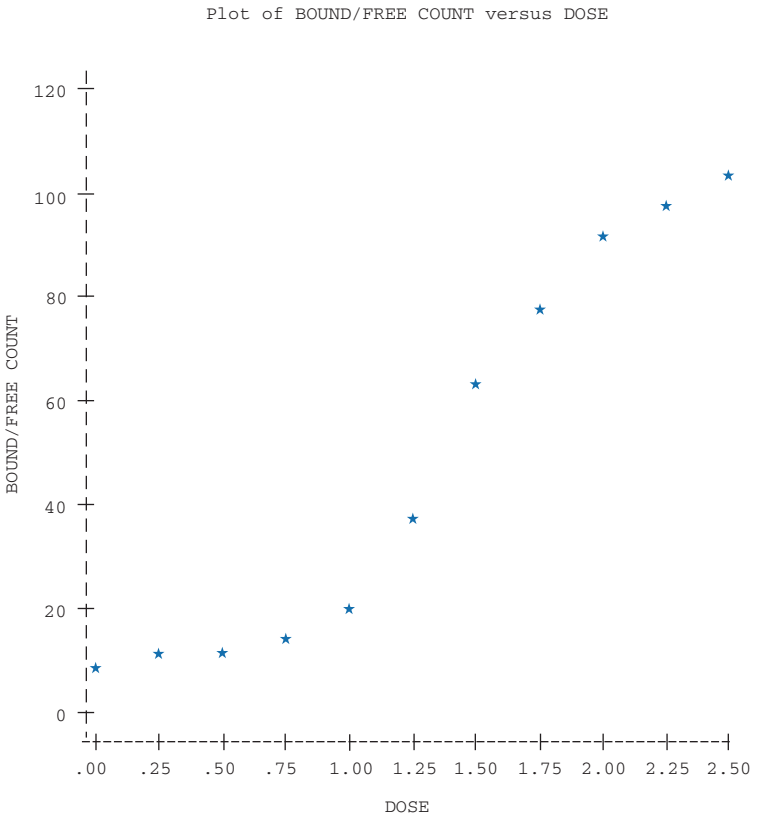
EXAMPLE 13.13

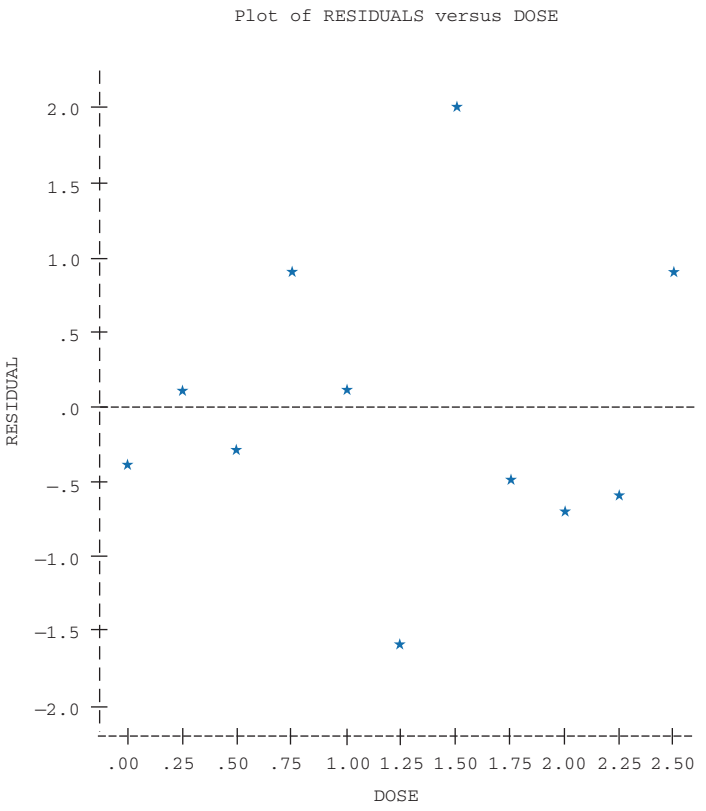
Use a nonlinear estimation program to fit the radioimmunoassay data to the model

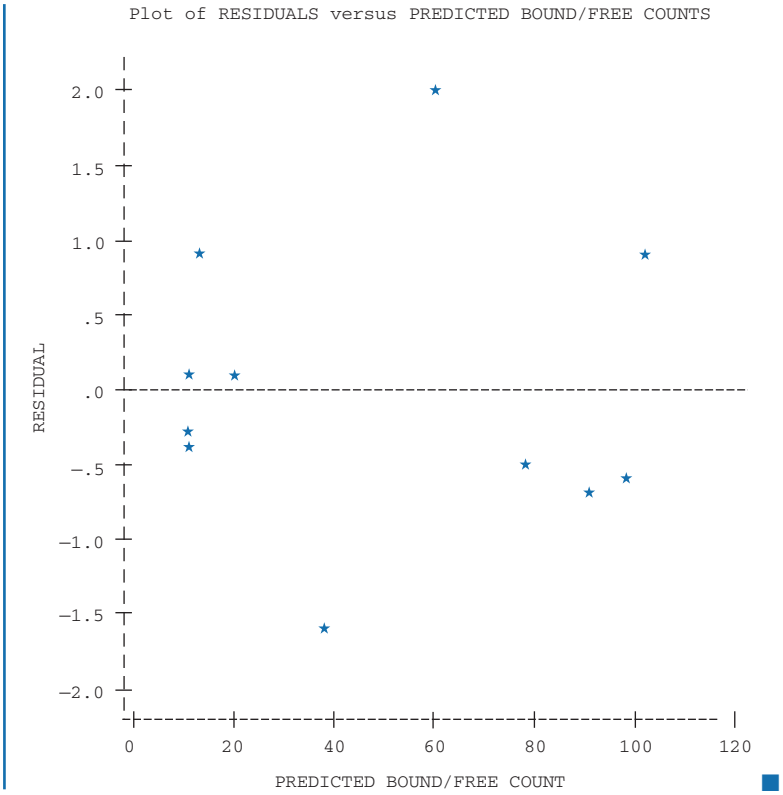
$$y = \frac{\beta_0 - \beta_3}{1 + (x/\beta_2)^{\beta_1}} + \beta_3$$

Solution SAS was used to fit this model to the sample data. As we can see from the residual plot, the nonlinear model provides a much better fit to the sample data than either the linear or the quadratic model.

```
NONLINEAR REGRESSION ANALYSIS

DATA LISTING

       BOUND/FREE
OBS    COUNT          DOSE

  1      9.900        0.00
  2     10.465        0.25
  3     10.312        0.50
  4     13.633        0.75
  5     20.784        1.00
  6     36.164        1.25
  7     62.045        1.50
  8     78.327        1.75
  9     90.307        2.00
 10     97.348        2.25
 11    102.686        2.50

Nonlinear Least Squares Summary Statistics   Dependent Variable COUNT

Source               DF  Sum of Squares   Mean Square

Regression            4    40390.959650  10097.739913
Residual              7        9.675063      1.382152
Uncorrected Total    11    40400.634713

(Corrected Total)    10    14673.985182

Parameter    Estimate    Asymptotic              Asymptotic 95%
                         Std. Error          Confidence Interval
                                             Lower         Upper
B0         10.3172019  0.6302496017     8.82688647   11.80751738
B1          5.3700868  0.2558475371     4.76509868    5.97507498
B2          1.4863334  0.0154121366     1.44988919    1.52277759
B3        107.3777343  1.7277534567   103.29221381  111.46325486

Asymptotic Correlation Matrix

Corr                B0               B1               B2               B3
-------------------------------------------------------------------------------
B0                   1      0.4317133357      0.1141723596      -0.255171767
B1        0.4317133357                 1      -0.514768068      -0.808689153
B2        0.1141723596      -0.514768068                 1      0.7939083509
B3        -0.255171767      -0.808689153      0.7939083509                 1

NOTE: Missing values were generated as a result of performing an operation on
      missing values. Each place is given by (number of times)
      AT (statement)/(line): (column) 4 AT 1/815:16
```







We can also use the fitted equation to predict y (count ratio) based on concentration.

13.4 Checking Model Assumptions (Step 3)

Now that we have identified possible independent variables (step 1) and considered the form of the multiple regression model (step 2), we should check whether the assumptions underlying the chosen model are valid. Recall that in Chapter 12 we indicated that the basic assumptions for a regression model of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

are as follows:

1. Zero expectation: $E(\varepsilon_i) = 0$ for all i.
2. Constant variance: $V(\varepsilon_i) = \sigma_\varepsilon^2$ for all i.
3. Normality: ε_i is normally distributed.
4. Independence: The ε_i are independent.

Note that because the assumptions for multiple regression are written in terms of the random errors ε_i, it would seem reasonable to check the assumptions by using the residuals $y_i - \hat{y}_i$, which are *estimates* of the ε_i.

The residuals are given by $e_i = y_i - \hat{y}_i$ and have mean 0 when the model has been correctly formulated and variances $\text{Var}(e_i) = \sigma_e^2(1 - h_{ii})$, where h_{ii} are the diagonal elements of the *hat matrix* $H = X(X'X)^{-1}X'$ and the X matrix is from the matrix formulation of the regression model as was discussed at the end of Chapter 12.

The first assumption, zero expectation, deals with model selection and whether additional independent variables need to be included in the model. If we have done our job in steps 1 and 2, assumption 1 should hold. The use of residual plots to check for inadequacy (lack of fit) of the model was discussed briefly in Chapter 11 and again in Section 13.3. If we have not done our job in steps 1 and 2, then a plot of the residuals should help detect this.

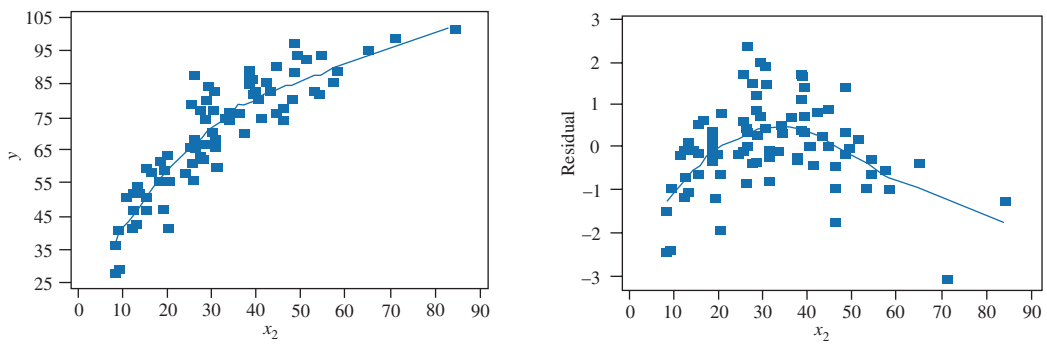
The residuals are standardized so that they have mean 0 and variance 1. The first choice of standardization is to divide the residual by $s_\varepsilon = \sqrt{\text{MSR}}$, where MSR is the mean square residual from the fitted model. This statistic is referred to as the *standardized* residual: e_i/s_ε. The problem with this standardization is that the standardized residuals do not have a variance equal to one. Thus, a more appropriate form for the standardization is to use the *studentized* residuals given by $d_i = e_i/s_\varepsilon\sqrt{1 - h_{ii}}$. The studentized residuals have a mean value of 0 and a variance of 1. The studentized residuals are available in most statistical software packages. Often, subtracting out the predictive part of the data reveals other structure more clearly. In particular, plotting the residuals from a first-order (linear terms only) model against each independent variable often reveals further structure in the data that can be used to improve the regression model.

One possibility is nonlinearity. We discussed nonlinearity and transformations earlier in the chapter. A noticeable curve in the residuals reflects a curved relation in the data, indicating that a different mathematical form for the regression equation would improve the predictive value of the model. A plot of residuals against each independent variable x often reveals this problem. A scatterplot smoother, such as LOWESS, can be useful in looking for curves in residual plots. For example, Figure 13.4 shows a scatterplot of y against x_2 and a residual plot against x_2. We think that the curved relation is more evident in the residual plot. The LOWESS curve helps considerably in both plots.

When nonlinearity is found, try transforming either independent or dependent variables. One standard method for doing this is to use (natural) logarithms of all variables except dummy variables. Such a model essentially estimates the *percentage* change in the dependent variable for a small percentage change in an independent variable, other independent variables held constant. Other useful transformations are logarithms of one or more independent variables only, square roots of independent variables, and inverses of the dependent variable or an independent variable. With a good computer package, a number of these transformations can be tested easily.

Assumption 2, the property of constant variance, can be examined using residual plots. One of the simplest residual plots for detecting nonconstant variance

FIGURE 13.4
y and residual plots showing curvature



is a plot of the residuals versus the predicted values, $\hat{y}_i$. Most of the available statistical software systems can provide these plots as part of the regression analysis.

EXAMPLE 13.14

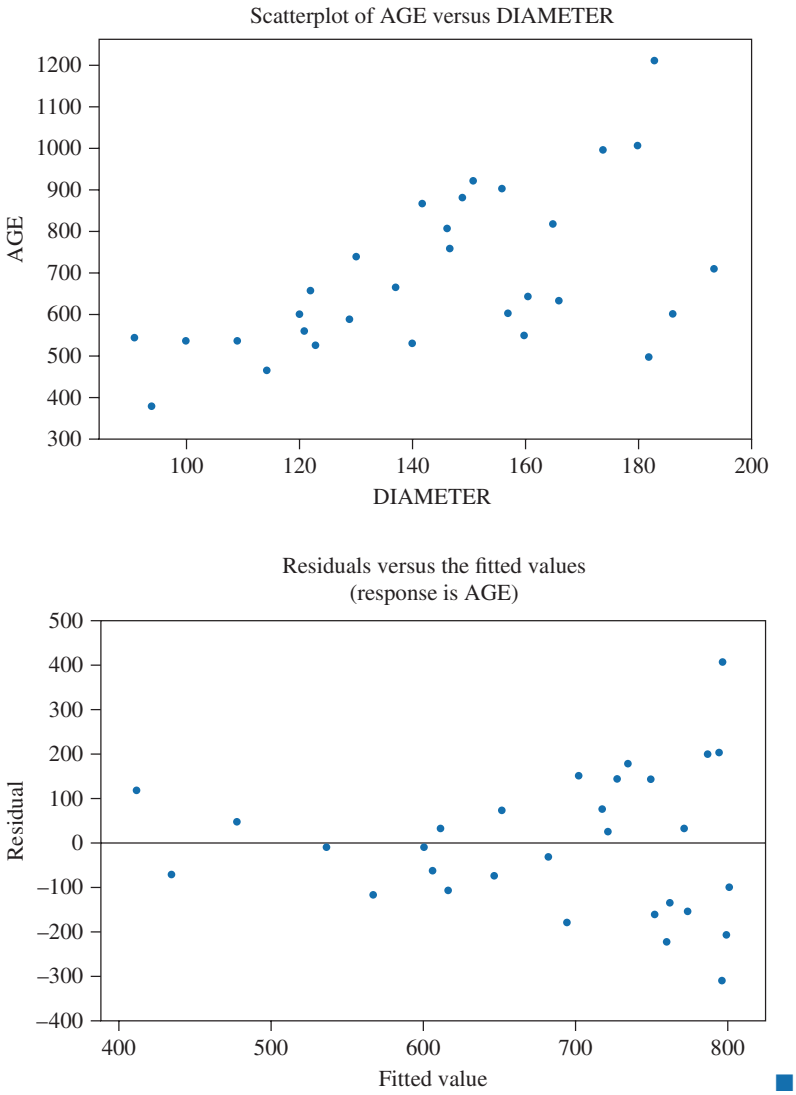
Forest scientists measured the diameters of 30 trees in a South American rain forest. The researchers then used carbon dating to determine the ages of the trees. The researchers were interested in determining if the diameter (D) of a tree in cm would provide an adequate prediction of the age (A) of the tree in years. The data are given in the Table 13.7.

TABLE 13.7
Tree age data

Tree	Diameter	Age	Tree	Diameter	Age	Tree	Diameter	Age
1	91	534	11	130	731	21	160	540
2	94	368	12	137	657	22	161	633
3	100	529	13	140	520	23	165	808
4	109	528	14	142	859	24	166	623
5	114	454	15	146	798	25	174	991
6	120	591	16	147	751	26	180	1,002
7	121	550	17	149	877	27	182	488
8	122	650	18	151	917	28	183	1,209
9	123	516	19	156	898	29	186	594
10	129	579	20	157	594	30	193	705

Solution The model $A = \beta_0 + \beta_1 D + \beta_2 D^2 + \varepsilon$ is fit to the data. As can be seen from the Minitab residual plot, the spread in the studentized residuals is generally increasing with the magnitudes of the predicted values of age, suggesting possible nonconstant variance of the studentized residuals. Also, because age is directly related to diameter via the regression model (i.e., age increases with diameter), the residuals are increasing with the magnitude of the values for diameter. This type of pattern in the residuals suggests that the variance of the ε_is (and hence the variance of the ages) is increasing with diameter. The accompanying plot of age versus diameter tends to support this observation.

```
Regression Analysis: AGE versus DIAMETER, DIA_SQ

The regression equation is
AGE = - 593 + 14.4 DIAMETER - 0.0374 DIA_SQ


Predictor      Coef    SE Coef      T      P
Constant    -592.5      732.2  -0.81  0.425
DIAMETER     14.44      10.50   1.38  0.180
DIA_SQ    -0.03741    0.03667  -1.02  0.317


S = 162.840  R-Sq = 33.4%  R-Sq(adj) = 28.4%


Analysis of Variance

Source          DF        SS      MS      F      P
Regression       2    358414  179207   6.76  0.004
Residual Error  27    715958   26517
Total           29   1074371
```



In some situations, there may be difficulties in readingresidual plots. In the book *Transformation and Weighting in Regression* (Carroll and Ruppert, 1988), it is pointed out that "the usual plots . . . are often sparse and difficult to interpret, particularly when the positive and negative residuals do not appear to exhibit the same general pattern. This difficulty is at least partially removed by plotting squared residuals . . . and thus visually doubling the sample size." There are several modifications that have been introduced for detecting heteroscedasticity of variance. These include plots of the absolute residuals, studentized residuals, and standardized residuals. The limitation of all graphical procedures is that they are all subjective and thus depend on the user's ability to differentiate "good" plots from "bad" plots. Attempts to remove this subjective nature of plot interpretation have resulted in several numerical measures of nonconstant variance. We will discuss one of these approaches, the Breusch–Pagan (BP) statistic.

The BP statistic tests the hypotheses H_0: homogeneous variances versus H_a: heterogeneous variances for the regression model. The BP statistic is discussed

in greater detail in *Applied Linear Regression Models* by Kutner, Nachtsheim, and Neter (2004). The BP procedure involves the following steps:

Step 1: Fit the regression model, $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$, to the data, and obtain the residuals, e_is, and the sum of squared residuals, SS(Residuals).

Step 2: Regress e_i^2 on the explanatory variables: Fit the model $e_i^2 = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \eta_i$, and obtain SS(Regression)*, the regression sum of squares from fitting the model with e_i^2 as the response variable.

Step 3: Compute the BP statistic:

$$\text{BP} = \frac{\text{SS(Regression)}^*/2}{(\text{SS(Residuals)}/n)^2}$$

where SS (Regression)* is the regression sum of squares from fitting the model with e_i^2 as the response variable and SS(Residuals) is the sum of square residuals from fitting the regression model with y as the response variable.

Step 4: Reject the null hypothesis of homogeneous variance if $\text{BP} > \chi^2_{\alpha,k}$, the upper α percentile from a squared distribution with degrees of freedom k.

Note: The residuals referred to in the BP procedure are the unstandardized residuals: $e_i = y_i - \hat{y}_i$.

Warning: The Breusch–Pagan test should be used only after it has been confirmed that the residuals have a normal distribution.

EXAMPLE 13.15

Refer to the data of Example 13.14, where the residual plots seemed to indicate a violation of the constant variance condition. Apply the Breusch–Pagan test to this data set, and determine if there is significant evidence of nonconstant variance.

Solution We will discuss methods for detecting whether or not the residuals appear to have a normal distribution at the end of this section. After that discussion, we will demonstrate in Example 13.17 that the residuals from the data in Example 13.14 appear to have a normal distribution. Thus, we can validly proceed to apply the BP test. Minitab output is given here.

```
Regression Analysis: AGE versus DIAMETER, DIA_SQ

Analysis of Variance

Source          DF        SS      MS      F      P
Regression       2    358414  179207   6.76  0.004
Residual Error  27    715958   26517
Total           29   1074371


Regression Analysis: RESID_SQ versus DIAMETER, DIA_SQ


Analysis of Variance

Source          DF            SS          MS      F      P
Regression       2   12341737513  6170868757   7.62  0.002
Residual Error  27   21859028491   809593648
Total           29   34200766004
```

From the first analysis of variance table, we obtain SS(Residual) = 715,958, and from the second analysis of variance table, we obtain SS(Regression)* = 12,341,737,513. We then compute

$$\text{BP} = \frac{\text{SS(Regression)}^*/2}{(\text{SS(Residuals)}/n)^2} = \frac{12{,}341{,}737{,}513/2}{(715{,}958/30)^2} = 10.83$$

The critical chi-squared value is $\chi^2_{\alpha,k} = \chi^2_{.05,2} = 5.99$. Because $\text{BP} = 10.83 > 5.99 = \chi^2_{.05,2}$ we reject H_0: homogenous variances and conclude that there is significant evidence that there is nonconstant variance in this situation. ■

weighted least squares

What are the consequences of having a nonconstant variance problem in a regression model? First, if the variance about the regression line is not constant, the least-squares estimates may not be as accurate as possible. A technique called **weighted least squares** (see Draper and Smith, 1998) will give more accuracy. Perhaps more important, however, the weighted least-squares technique improves the statistical tests (F and t tests) on model parameters and the interval estimates for parameters because they are, in general, based on smaller standard errors.

The more serious pitfall involved with inferences in the presence of nonconstant variance seems to be for estimates $E(y)$ and predictions of y. For these inferences, the point estimate y is sound, but the width of the interval may be too large or too small depending on whether we're predicting in a low- or high-variance section of the experimental region.

The best remedy for nonconstant variance is to use weighted least squares. We will not cover this technique in the text. However, when the nonconstant variance possesses a pattern related to y, a reexpression (transformation) of y may resolve the problem. Several transformations for y were discussed in Chapter 11; ones that help to stabilize the variance when there is a pattern to the nonconstant variance were discussed in Chapter 8 for the analysis of variance. They can also be applied in certain regression situations.

Box–Cox

An excellent discussion of transformations is given in the book *Introduction to Regression Modeling* by Abraham and Ledolter (2006). A special class of transformations is called **Box–Cox** transformations. The general form of the Box–Cox transformation is

$$g(y_i) = (y_i^\lambda - 1)/\lambda$$

where λ is a constant to be determined from the data. From the form of $g(y_i)$, we can observe the following special cases:

- If $\lambda = 1$, then no transformation is needed. The original data should be modeled.
- If $\lambda = 2$, then the Box–Cox transformation is the square of the original response variable, and y_i^2 should be modeled.
- If $\lambda = -1$, then the Box–Cox transformation is the reciprocal of the original response variable, and $1/y_i$ should be modeled.
- If $\lambda = 1/2$, then the Box–Cox transformation is the reciprocal of the original response variable, and $\sqrt{y_i}$ should be modeled.
- If $\lambda = 0$, then in the limit as λ converges to 0, the Box–Cox transformation is the natural logarithm of the original response variable, and $\log(y_i)$ should be modeled.
- If $\lambda = -1/2$, then the Box–Cox transformation is the reciprocal of the square root of the original response variable, and $1/\sqrt{y_i}$ should be modeled.

In the article "An Analysis of Transformation," Box and Cox (1964) describe a process to obtain a sample estimate of λ. The steps in their process are as given here. Define $y^{(\lambda)}$ by

$$y_i^{(\lambda)} = \frac{(y_i^{\lambda} - 1)}{\lambda \bar{y}_g^{\lambda - 1}}$$

where $\bar{y}_g = [\prod_{i=1}^{n} y_i]^{1/n}$ is the geometric mean of the values of the response variable, y_i. If $\lambda = 0$, then $y^{(\lambda)}$ would be undefined. Thus, when $\lambda = 0$, we take its limiting value:

$$y^{(\lambda = 0)} = \lim_{\lambda \to 0} y_i^{(\lambda)} = \bar{y}_g \log(y_i)$$

where $\log(y_i)$ is the natural logarithm. To obtain an estimate of λ, follow these steps:

Step 1: Select a grid of values for λ:

$$\lambda = -2, -1.75, -1.5, -1.25, -1.0, -.75, -.5, -.25, 0, .25, .50, .75, 1.0, 1.25, 1.5, 1.75, 2$$

Step 2: For each value of λ in the grid, regress $y^{(\lambda)}$ on the k explanatory variables, and obtain the SS(Residual) from the fitted model.

Step 3: Take as your value for λ that value of λ having the smallest value of SS(Residual).

EXAMPLE 13.16

Refer to Example 13.15, where we detected a violation of the constant variance condition. Determine the Box–Cox transformation for this data set. Regress the transformed variable, and determine if there is an improvement of the model fit and a reduction in the heterogeneity of the variances.

Solution Table 13.8 gives the values of MS(Residual) for the various values of λ.

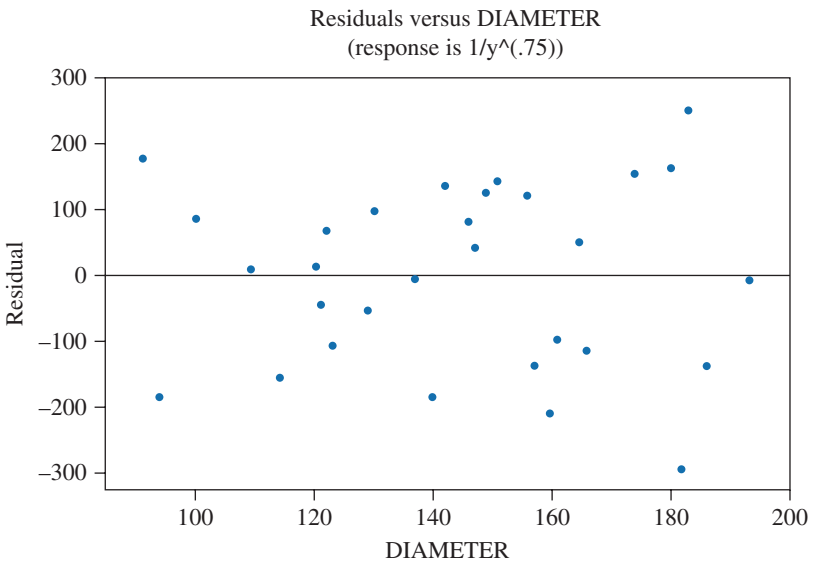
TABLE 13.8
MS(Residual) as a function of λ

λ	MS(Residual)	λ	MS(Residual)
2.00	1,039,501	−.25	556,310
1.75	934,661	−.50	546,340
1.50	847,632	−.75	543,015
1.25	775,501	−1.00	546,517
1.00	715,958	−1.25	557,276
.75	667,182	−1.50	575,994
.50	627,761	−1.75	603,686
.25	596,619	−2.00	641,736
0	572,976		

From Table 13.8, the value of λ that yields the smallest value for MS(Residual) is $\lambda = -.75$. To determine if the transformation $y_i^{-.75} = 1/y_i^{.75}$ yields an improved fit, the model $1/y^{.75} = \beta_0 + \beta_1 Diameter + \beta_2 Diameter^2 + \varepsilon$ was fit to the data. The Minitab package produced the following output.

```
Regression Analysis: 1/y^(.75) versus DIAMETER, DIA_SQ

The regression equation is
1/y^(.75) = 111612 + 21.3 DIAMETER - 0.0619 DIA_SQ


Predictor      Coef    SE Coef       T      P
Constant     111612        638  175.03  0.000
DIAMETER     21.271      9.142    2.33  0.028
DIA_SQ     -0.06187    0.03194   -1.94  0.063


S = 141.816  R-Sq = 41.4%  R-Sq(adj) = 37.0%


Analysis of Variance

Source          DF        SS      MS      F      P
Regression       2    383229  191615   9.53  0.001
Residual Error  27    543015   20112
Total           29    926244
```

The plot of residuals versus diameter is given below.

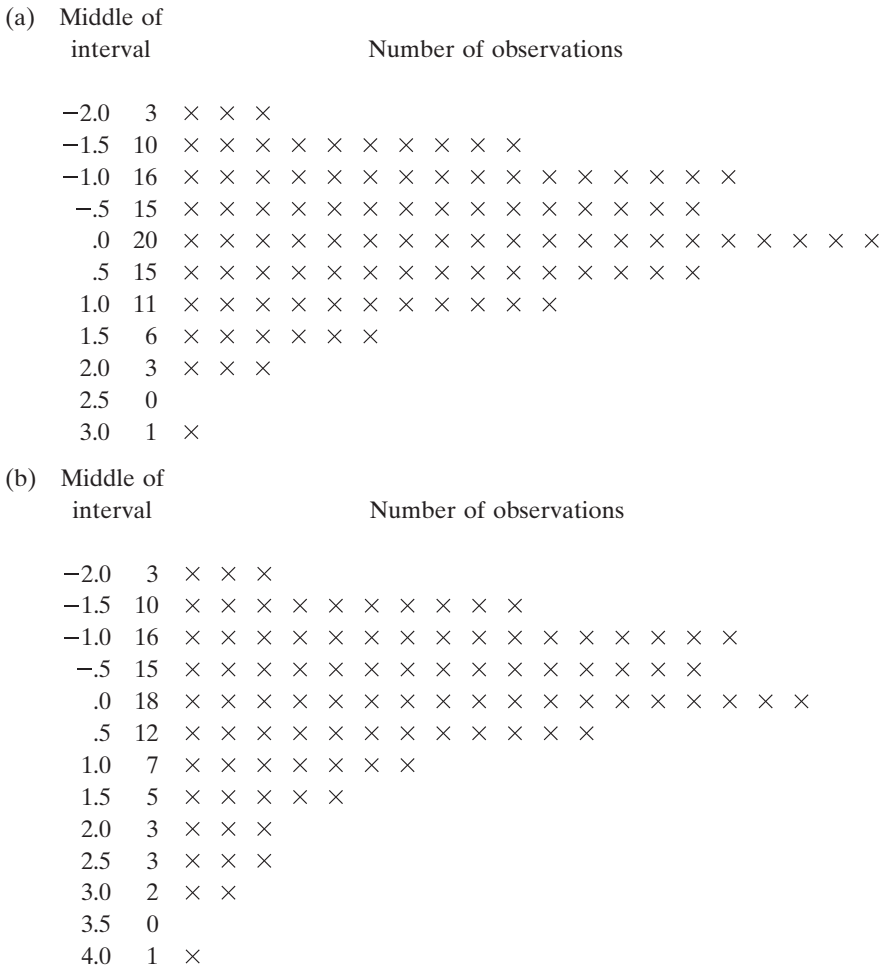


From the residual plot, it would appear that the nonconstant variance pattern that was present in the residuals when using the model involving the untransformed age variable has been greatly reduced using the transformed age variable. The BP test was computed for the transformed data, yielding the following results:

$$\text{BP} = \frac{\text{SS(Regression)}^*/2}{(\text{SS(Residuals)}/n)^2} = \frac{2{,}186{,}828{,}520/2}{(543{,}015/30)^2} = 3.34$$

The critical chi-squared value is $\chi^2_{\alpha,k} = \chi^2_{.05,2} = 5.99$. Because $\text{BP} = 3.34 < 5.99 = \chi^2_{.05,2}$, we fail to reject H_0: homogenous variances and conclude that there is not significant evidence of nonconstant variance in this situation. The Box–Cox transformation has eliminated the violation of the constant variance condition. Also, the value of R^2 has increased from 33.4% from the model using the original y values to 41.1% for the model fit using the Box–Cox transformation. ■

FIGURE 13.5
Top: residuals centered on zero; bottom: residuals skewed to right



The third assumption for multiple regression is that of normality of the ε_i. Skewness and/or outliers are examples of forms of nonnormality that may be detected through the use of certain scatterplots and residual plots.

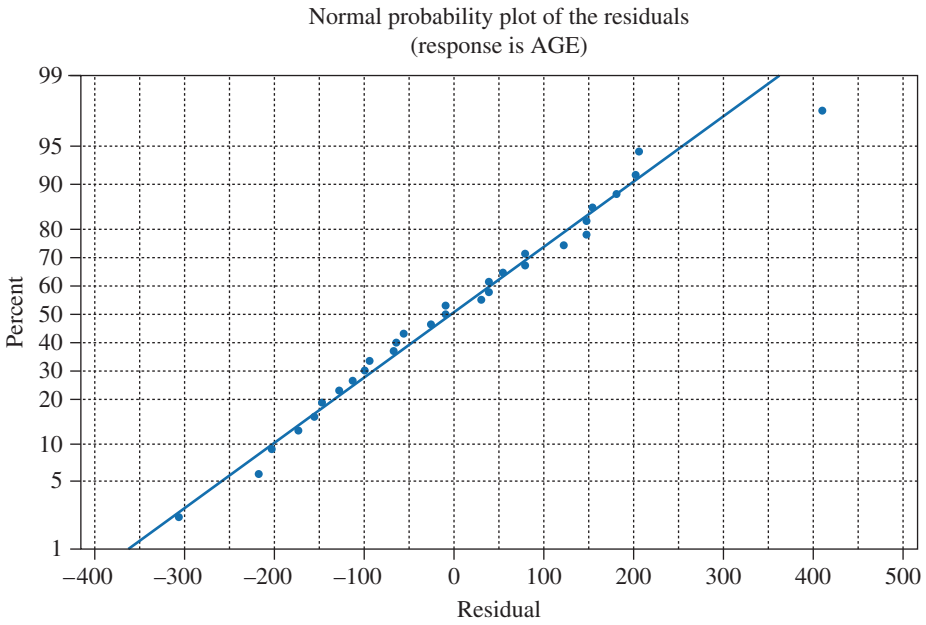
A plot of the residuals in the form of a histogram or a stem-and-leaf plot will help to detect skewness. By assumption, the ε_i are normally distributed with mean 0. If a histogram of the residuals is not symmetrical about 0, some skewness is present. For example, the residual plot in Figure 13.5(a) is symmetrical on 0 and suggests no skewness. In contrast, the residual plot in Figure 13.5(b) is skewed to the right.

probability plot

Another way to detect nonnormality is through the use of a normal **probability plot** of the residuals, as was discussed in Chapter 4. The idea behind the plot is that if the residuals are normally distributed, the normal probability plot will be approximately a straight line. Most computer packages in statistics offer an option to obtain normal probability plots. We'll use them when needed to do our plots.

EXAMPLE 13.17

Refer to the data in Example 13.14. Use the normal probability plot following to determine whether there is evidence that the distribution of the residuals has a nonnormal distribution.



Solution The plotted points in the normal probability plot fall very close to the straight line. Thus, we can be reasonably assured that the residuals have a normal distribution. ■

The presence of one or more outliers is perhaps a more subtle form of nonnormality that may be detected by using a scatterplot and one or more residual plots. An outlier is a data point that falls away from the rest of the data. Recall from Chapter 11 that we must be concerned about the leverage (x outlier) and influence (both x and y outlier) properties of a point. A high influence point may seriously distort the regression equation. In addition, some outliers may signal a need for taking some action. For example, if a regression analysis indicates that the price of a particular parcel of land is very much lower than predicted, that parcel may be an excellent purchase. A sales office that has far better results than a regression model predicts may have employees who are doing outstanding work that can be copied. Conversely, a sales office that has far poorer results than the model predicts may have problems. Sometimes it is possible to isolate the reason for the outlier; other times it is not. An outlier may arise because an error was made in recording the data or in entering it into a computer or because the observation is obtained under different conditions from the other observations. If such a reason can be found, the data entry can be corrected or the point omitted from the analysis. If there is no identifiable reason to correct or omit the point, run the regression both with and without it to see which results are sensitive to that point. No matter what the source or reason for outliers, if they go undetected, they can cause serious distortions in a regression equation.

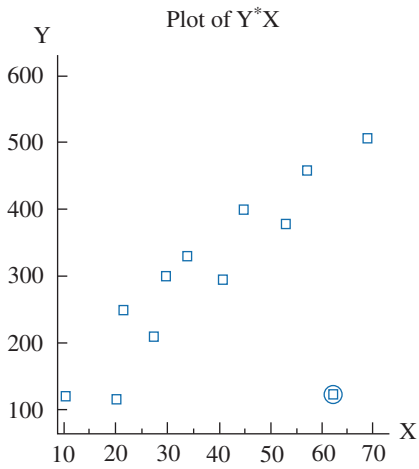
For the linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, a scatterplot of y versus x will help detect the presence of an outlier. This is shown in Table 13.9 and Figure 13.6. It certainly appears that the circled data point is an outlier. Computer output for a linear fit to the data of Table 13.9 is shown here, along with a residual plot and a normal probability plot. Again, the data point corresponding to the suspected outlier (62, 125) is circled in each plot. The Minitab program produced the following analysis.

TABLE 13.9
Listing of data

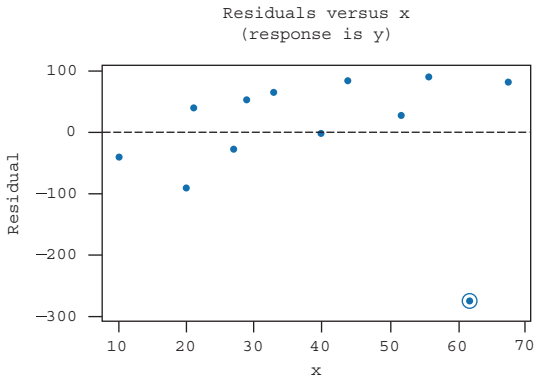
Obs	*x*	*y*
1	10	120
2	20	115
3	21	250
4	27	210
5	29	300
6	33	330
7	40	295
8	44	400
9	52	380
10	56	460
11	62	125
12	68	510

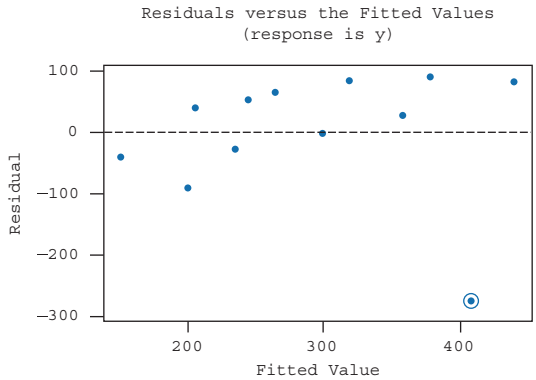
$N = 12$

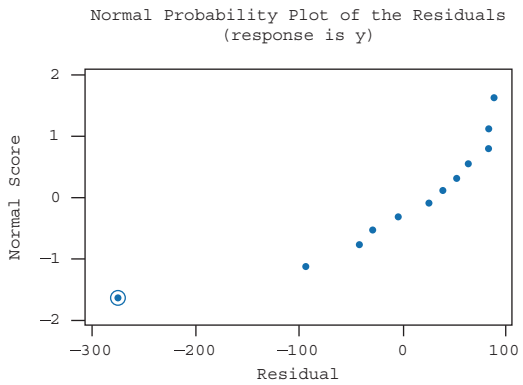
FIGURE 13.6
Scatterplot of the data in Table 13.9




```
Regression Analysis: y versus x


The regression equation is
y = 114 + 4.59 x

Predictor        Coef     SE Coef          T         P
Constant       114.36       75.53       1.51     0.161
x               4.595       1.787       2.57     0.028

s = 108.1       R-Sq = 39.8%     R-Sq(adj) = 33.8%

Analysis of Variance

Source            DF          SS          MS        F        P
Regression         1       77201       77201     6.61    0.028
Residual Error    10      116755       11676
Total             11      193956

Obs          x          y         Fit       SE Fit     Residual   Standardized
                                                                    Residual
  1       10.0      120.0       160.3        59.7        -40.3       -0.45
  2       20.0      115.0       206.2        45.4        -91.2       -0.93
  3       21.0      250.0       210.8        44.2         39.2        0.40
  4       27.0      210.0       238.4        37.4        -28.4       -0.28
  5       29.0      300.0       247.6        35.5         52.4        0.51
  6       33.0      330.0       266.0        32.7         64.0        0.62
  7       40.0      295.0       298.1        31.3         -3.1       -0.03
  8       44.0      400.0       316.5        32.7         83.5        0.81
  9       52.0      380.0       353.3        39.4         26.7        0.27
 10       56.0      460.0       371.7        44.2         88.3        0.90
 11       62.0      125.0       399.2        52.3       -274.2       -2.90R
 12       68.0      510.0       426.8        61.2         83.2        0.93

R denotes an observation with a large standardized residual
```







This data set helps to illustrate one of the problems in trying to identify outliers. Sometimes a single plot is not sufficient. For this example, the scatterplot and the probability plot clearly identify the outlier, whereas the residual plot is less conclusive because the outlier adversely affects the linear fit to the data by pulling the fitted line toward the outlier. This makes some of the other residuals larger than they should be. The message is clear: *Don't jump to conclusions without examining the data in several different ways.* The problem becomes even more difficult with multiple regression, where simple scatterplots are not possible.

When we have multiple explanatory variables, it is possible that data points having high leverage and/or high influence may not be detected by just plotting the data. There are a number of diagnostics that are outputted by most statistical software packages. Two of the most commonly used statistics are h_{ii}, the diagonal elements of the Hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and Cook's D statistic. The values of h_{ii} are used to determine if the ith observation $(y_i, x_{1i}, x_{2i}, \ldots, x_{ki})$ has high leverage. If $h_{ii} > 2(k+1)/n$, then the ith observation is considered high leverage in the fit of the regression model. Such an ith observation needs to be identified and then given a careful examination to determine if the values of the explanatory variables in that observation have been misrecorded or if they are much different than those of the remaining $n - 1$ observations. A high leverage value may or may not have high influence.

Cook's D statistic attempts to identify observations that have high influence by measuring how the deletion of an observation affects the parameter estimates. Let $\hat{\boldsymbol{\beta}}$ be the estimates of the regression coefficients obtained from the full data set and $\hat{\boldsymbol{\beta}}_{(i)}$ be the vector of estimates of the regression coefficients obtained from the data set in which the ith observation has been deleted. Cook's D statistic measures the difference between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$. How large must Cook's D be for an observation to need to be examined? There is no trigger value as there was in the case of h_{ii}. The values of Cook's D should be used to compare the n observations for influence. Select those observations having the largest value for D. In the literature, it is often recommended that if an observation has a value of D greater than 1, then this observation demands examination.

EXAMPLE 13.18

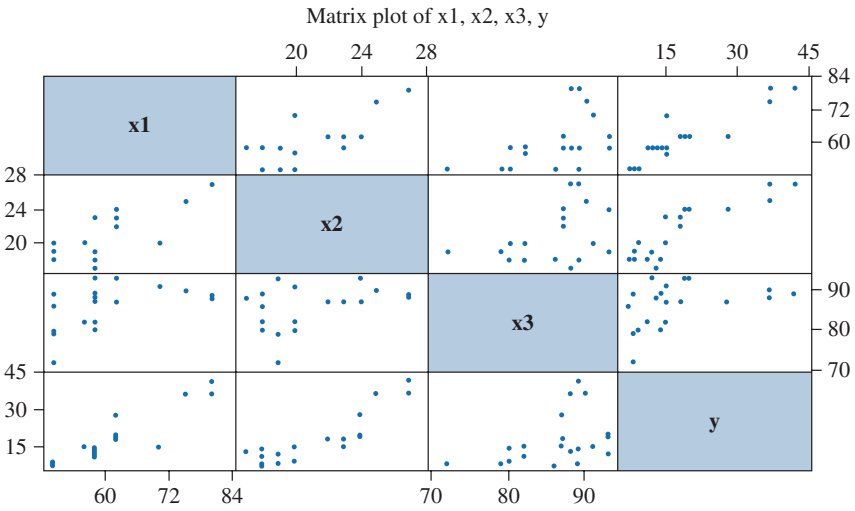
An example that has often been used to illustrate the detection of high leverage and high influence is the *Brownlee's stack-loss* data. The data given below were obtained from 21 days of operation of a plant for the oxidation of ammonia to nitric acid and are presented in **Statistical Theory and Methodology in Science and Engineering *(Brownlee, 1965).*** The dependent variable is 10 times the percentage of the ingoing ammonia to the plant that escapes unabsorbed. The explanatory variables are x_1 = airflow, x_2 = cooling water inlet temperature, and x_3 = acid concentration. The data are given in Table 13.10.

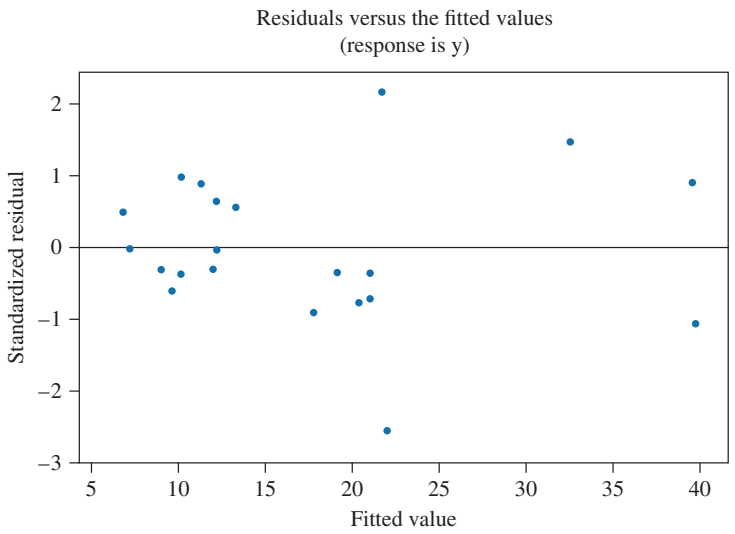
TABLE 13.10
Stack-loss data

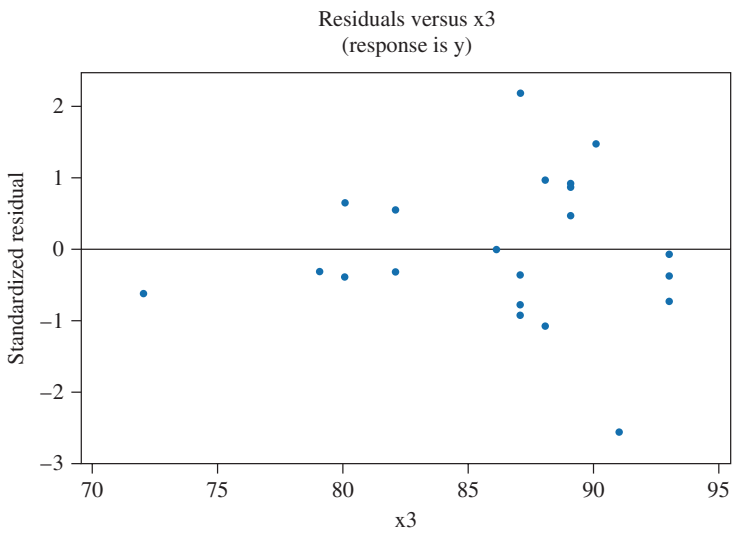
Case	x_1	x_2	x_3	y	Case	x_1	x_2	x_3	y
1	80	27	89	42	12	58	17	88	13
2	80	27	88	37	13	58	18	82	11
3	75	25	90	37	14	58	19	93	12
4	62	24	87	28	15	50	18	89	8
5	62	22	87	18	16	50	18	86	7
6	62	23	87	18	17	50	19	72	8
7	62	24	93	19	18	50	19	79	8
8	62	24	93	20	19	50	20	80	9
9	58	23	87	15	20	56	20	82	15
10	58	18	80	14	21	70	20	91	15
11	58	18	89	14					

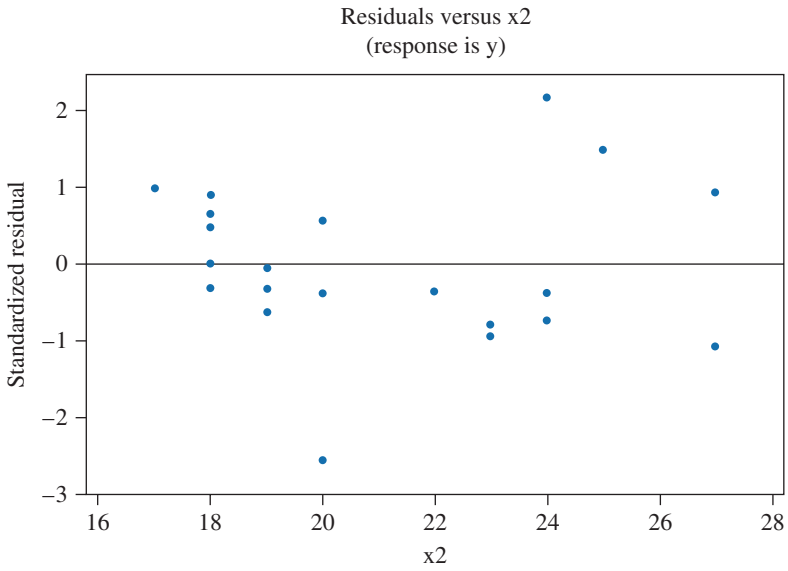
The model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \varepsilon$ was fit to the data, yielding the following Minitab output, scatterplot matrix, and residual plots.

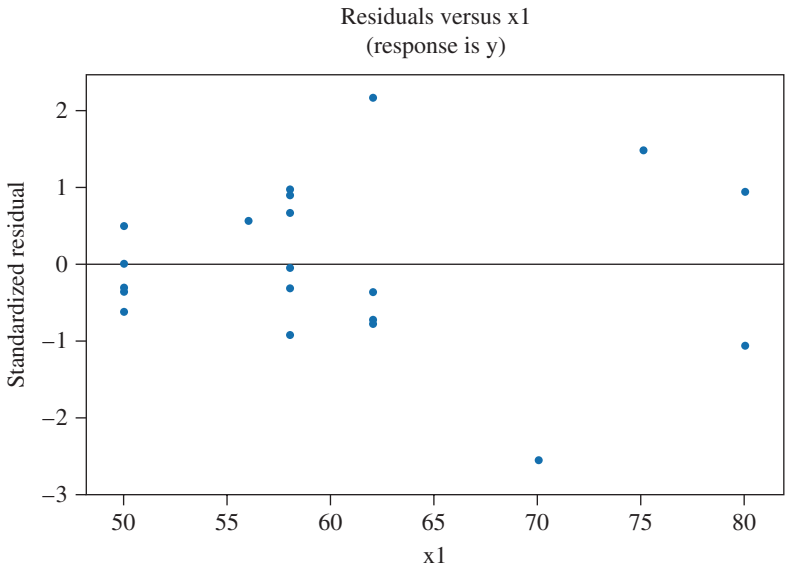
```
Regression Analysis: y versus x1, x2, x3, x1_sq

The regression equation is
y = - 16.4 - 0.17 x1 + 1.26 x2 - 0.093 x3 + 0.00678 x1_sq

Predictor      Coef    SE Coef        T      P    VIF
Constant    -16.35      33.29    -0.49  0.630
x1          -0.165      1.168    -0.14  0.889  212.6
x2           1.2613     0.3754    3.36  0.004    2.6
x3          -0.0934     0.1762   -0.53  0.603    1.7
x1_sq     0.006784   0.008933    0.76  0.459  207.2

S = 3.28452  R-Sq = 91.7%  R-Sq(adj) = 89.6%

Analysis of Variance

Source           DF         SS       MS      F       P
Regression        4    1896.63   474.16  43.95  0.000
Residual Error   16     172.61    10.79
  Lack of Fit    15     172.11    11.47  22.95  0.163
  Pure Error      1       0.50     0.50
Total            20    2069.24

Unusual Observations

Obs    x1       y     Fit  SE Fit  Residual  St Resid
  4  62.0  28.000  21.623   1.478     6.377      2.17R
 21  70.0  15.000  22.046   1.770    -7.046     -2.55R

R denotes an observation with a large standardized residual.


Case  x1  x2  x3   y      SRES1       HI1      COOK1
  1   80  27  89  42    0.95685  0.409572   0.127022
  2   80  27  88  37   -1.06253  0.410937   0.157516
  3   75  25  90  37    1.49660  0.176019   0.095694
  4   62  24  87  28    2.17418  0.202615   0.240228
  5   62  22  87  18   -0.35564  0.112237   0.003198
  6   62  23  87  18   -0.77741  0.144365   0.020394
  7   62  24  93  19   -0.71871  0.236391   0.031981
  8   62  24  93  20   -0.37030  0.236391   0.008490
  9   58  23  87  15   -0.92079  0.163108   0.033049
 10   58  18  80  14    0.66822  0.261592   0.031637
 11   58  18  89  14    0.90379  0.156344   0.030274
 12   58  17  88  13    0.99738  0.219303   0.055887
 13   58  18  82  11   -0.31521  0.197933   0.004904
 14   58  19  93  12   -0.05511  0.207790   0.000159
 15   50  18  89   8    0.49077  0.383454   0.029959
 16   50  18  86   7   -0.00516  0.266996   0.000002
 17   50  19  72   8   -0.62911  0.412771   0.055639
 18   50  19  79   8   -0.31581  0.196788   0.004887
 19   50  20  80   9   -0.37705  0.214605   0.007769
 20   56  20  82  15    0.56698  0.100353   0.007172
 21   70  20  91  15   -2.54670  0.290440   0.530949
```











An examination of the scatterplot matrix and residual plots reveals a few observations that may need further investigation. Cases 4 and 21 have large-in-magnitude standardized residuals. Cases 1 and 2 are both at the outer edge of values for all three explanatory variables and may have high leverage. The table of values for the leverage values h_{ii} and D values reveals that cases 4 and 21 have standardized residuals of 2.174 and −2.547, respectively. Both of these values would be considered large. The values of h_{ii} for cases 1, 2, 4, and 21 are .4095, .4109, .2026, and .2904, respectively. Using the criterion $h_{ii} > 2(k + 1)/n = 2(5)/21 = .476$, none of these values would indicate a concern for high leverage. The case having the highest leverage value was case 17: $h_{ii} = .4128 < .476$, and, hence, it should not be considered of high leverage. It may be noted that case 17 had the lowest values for x_1 and x_3 and hence placed itself in a corner of the observation space. Next, we will examine the values of Cook's D. The cases with largest values are cases 4 and 21.

Because neither of these cases had high leverage, their high values of D are due to their large standardized residuals. To evaluate the impact of these two cases, the regression models were rerun three times first with case 4 deleted, then with case 21 deleted, and finally with both cases deleted. The results are summarized in Table 13.11.

TABLE 13.11
Impact of outliers on parameter estimates

Parameter Estimate	All Data	w/o Case 4	w/o Case 21	w/o Cases 4, 21
$\hat{\beta}_0$	−16.35	5.84	−27.28	−4.19
$\hat{\beta}_1$	−.165	−.871	.2745	−.4557
$\hat{\beta}_2$	1.2613	.9762	.8018	.4772
$\hat{\beta}_3$	−.0934	−.0469	−.0672	−.0166
$\hat{\beta}_4$	.00678	.0127	.00471	.0109
Statistics				
R^2	91.7%	93.8%	95.0%	97.7%
MSE	10.79	8.11	6.84	3.22

From Table 13.11, it is obvious that both cases 4 and 21 have a strong influence on the fit of the regression model. There is a large change in the estimation regression coefficients, an increase in R^2, and a decrease in MSE when either or both of the cases are removed from the data set. The researchers would next have to carefully examine the data associated with these two cases and the conditions under which the data were collected. A decision to delete one or both of the cases would then be made. However, if cases are removed from the data set, it is always good practice to include in any papers or reports a listing of these cases and an explanation of why they were deleted. ■

If you detect outliers, what should you do with them? Of course, recording or transcribing errors should simply be corrected. Sometimes an outlier obviously comes from a different population than the other data points. For example, a Fortune 500 conglomerate firm doesn't belong in a study of small manufacturers. In such situations, the outliers can reasonably be omitted from the data. Unless a compelling reason can be found, throwing out a data point is inappropriate.

The final assumption is that the ε_i are statistically independent and hence uncorrelated. When the time sequence of the observations is known, as is the case with **time series** data, where observations are taken at successive points in time, it is possible to construct a plot of the residuals versus time to observe where the residuals are **serially correlated.** If, for example, there is a positive serial correlation, adjacent residuals (in time) tend to be similar; negative serial correlation implies that adjacent residuals are dissimilar. These patterns of positive and negative serial correlation are displayed in Figures 13.7(a) and 13.7(b), respectively. Figure 13.7(c) shows a residual plot with no apparent serial correlation.

time series

serial correlation

A formal statistical test for serial correlation is based on the *Durbin–Watson statistic.* Let e_t denote the residual at time t and n the total number of time points. Then the **Durbin–Watson test statistic** is

Durbin–Watson test statistic

$$d = \frac{\sum_{t=1}^{n-1}(e_{t+1} - e_t)^2}{\sum_t e_t^2}$$

The logic behind this statistic is as follows: If there is a positive serial correlation, then successive residuals will be similar and their squared difference, $(e_{t+1} - e_t)^2$,

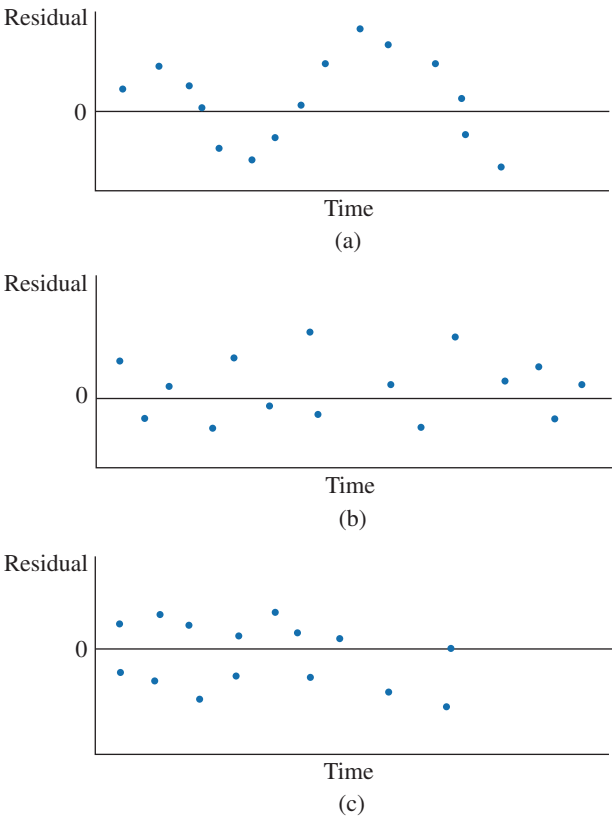


FIGURE 13.7
(a) Positive serial correlation. (b) Negative serial correlation. (c) No apparent serial correlation.

will tend to be smaller than it would be if the residuals were uncorrelated. Similarly, if there is a negative serial correlation among the residuals, the squared difference of successive residuals will tend to be larger than when no correlation exists.

positive and negative serial correlation

When there is no serial correlation, the expected value of the Durbin–Watson test statistic d is approximately 2.0; **positive serial correlation** makes $d < 2.0$ and **negative serial correlation** makes $d > 2.0$. Although critical values of d have been tabulated by Durbin and Watson (1951), values of d less than approximately 1.5 (or greater than approximately 2.5) lead one to suspect positive (or negative) serial correlation.

EXAMPLE 13.19

Sample data corresponding to retail sales for a particular line of personalized computers by month are shown in Table 13.12.

TABLE 13.12
Sales data

Month, x	Sales, y (millions of dollars)	Month, x	Sales, y (millions of dollars)
1	6.0	8	8.5
2	6.3	9	9.0
3	6.1	10	8.7
4	6.8	11	7.9
5	7.5	12	8.2
6	8.0	13	8.4
7	8.1	14	9.0

Plot the data. Also plot the residuals by time based on a linear regression equation. Does there appear to be serial correlation?

Solution It is clear from the scatterplot of the sample data and from the residual plot of the linear regression that serial correlation is present in the data.

OBS	MONTH SALE	COMPUTER SALES (MILLIONS OF DOLLARS)
1	1	6.0
2	2	6.3
3	3	6.1
4	4	6.8
5	5	7.5
6	6	8.0
7	7	8.1
8	8	8.5
9	9	9.0
10	10	8.7
11	11	7.9
12	12	8.2
13	13	8.4
14	14	9.0

Dependent Variables: Y SALES (MILLIONS OF DOLLARS)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	10.57540	10.57540	34.302	0.0001
Error	12	3.69960	0.30830		
C Total	13	14.27500			

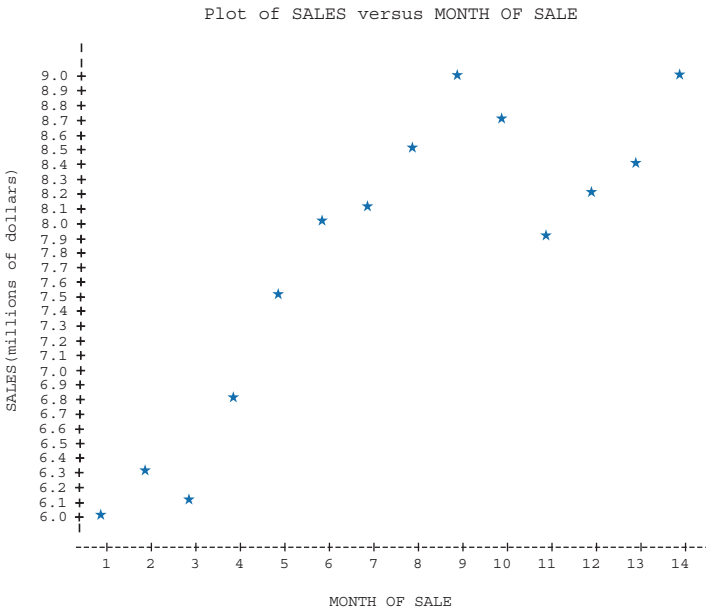
Root MSE	0.55525	R-square	0.7408
Dep Mean	7.75000	Adj R-sq	0.7192
C.V.	7.16449		

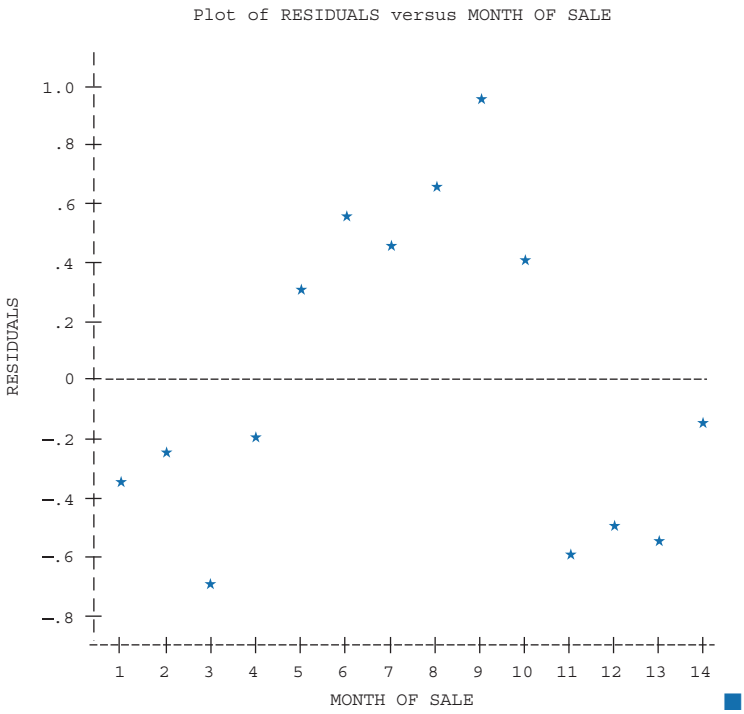
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	6.132967	0.31344787	19.566	0.0001
X	1	0.215604	0.03681259	5.857	0.0001

Variable	DF	Variable Label
INTERCEP	1	Intercept
X	1	MONTH

Durbin-Watson D	0.625
(For Number of Obs.)	14
1st Order Autocorrelation	0.668





EXAMPLE 13.20

Determine the value of the Durbin–Watson statistic for the data of Example 13.19. Does it confirm the impressions you obtained from the plots?

Solution Based on the output of Example 13.19, we find $d = .625$. Because this value is much less than 1.5, we have evidence of positive serial correlation; the residual plot bears this out. ■

If serial correlation is suspected, then the proposed multiple regression model is inappropriate, and some alternative must be sought. A study of the many approaches to analyzing time series data where the errors are not independent can consume many years; hence, we cannot expect to solve many of these problems within the confines of this text. We will, however, suggest a simplified regression approach, based on *first differences,* which may alleviate the problem.

Regression based on first differences is simple to use and, as might be expected, is only a crude approach to the problem of serial correlation. For a simple linear regression of y on x, we compute the differences $y_t - y_{t-1}$ and $x_t - x_{t-1}$. A regression of the $n - 1$ y differences on the corresponding $n - 1$ x differences may eliminate the serial correlation. If not, you should consult someone more familiar with analyzing time series data.

The residual plots that we have discussed can be useful in diagnosing problems in fitting regression models to data. Unfortunately however, they, too, can be misleading because the residuals are subject to random variation. Some researchers have suggested that it is better to use ''standardized'' residuals to detect problems with a fitted regression model.

If the software package you use works with standardized residuals, you can replace plots of the ordinary residuals with plots of the standardized residuals to perform the diagnostic evaluation of the fit of a regression model. In theory, these standardized residuals have a mean of 0 and a standard deviation of 1. Large residuals would be ones with an absolute value of, say, 3 or more.

13.5 RESEARCH STUDY: Construction Costs for Nuclear Power Plants

One of the major issues confronting power companies in seeking alternatives to fossil fuels the need to forecast the costs of constructing nuclear power plants. The data documenting the construction costs of 32 light water reactor (LWR) nuclear power plants, constructed in late 1960s and early 1970s, along with information on the construction of the plants and specific characteristics of each power plant are presented in Table 13.13. The research goal is to determine which of the explanatory variables are most strongly related to the capital cost of the plant. If a reasonable model can be produced from these data, then the construction costs of new plants meeting specified characteristics can be predicted. Because of the resistance of the public and politicians to the construction of nuclear power plants, there is only a limited amount of data associated with new construction. The data set provided by ***Cox and Snell (1981)*** has only $n = 32$ plants along with 10 explanatory variables. The book *Introduction to Regression Modeling* (Abraham and Ledolter, 2006) provides a detailed analysis of this data set. We will document some of the steps needed to build a model and then assess its usefulness in predicting the cost of construction of specific types of nuclear power plants. This is a relatively small data set ($n = 32$) especially considering the large number of explanatory variables ($k = 10$).

TABLE 13.13
Power plant construction costs data

Plant	C	D	T1	T2	S	PR	NE	CT	BW	N	PT
1	460.05	68.58	14	46	687	0	1	0	0	14	0
2	452.99	67.33	10	73	1,065	0	0	1	0	1	0
3	443.22	67.33	10	85	1,065	1	0	1	0	1	0
4	652.32	68	11	67	1,065	0	1	1	0	12	0
5	642.23	68	11	78	1,065	1	1	1	0	12	0
6	345.39	67.92	13	51	514	0	1	1	0	3	0
7	272.37	68.17	12	50	822	0	0	0	0	5	0
8	317.21	68.42	14	59	457	0	0	0	0	1	0
9	457.12	68.42	15	55	822	1	0	0	0	5	0
10	690.19	68.33	12	71	792	0	1	1	1	2	0
11	350.63	68.58	12	64	560	0	0	0	0	3	0
12	402.59	68.75	13	47	790	0	1	0	0	6	0
13	412.18	68.42	15	62	530	0	0	1	0	2	0
14	495.58	68.92	17	52	1,050	0	0	0	0	7	0
15	394.36	68.92	13	65	850	0	0	0	1	16	0
16	423.32	68.42	11	67	778	0	0	0	0	3	0
17	712.27	69.5	18	60	845	0	1	0	0	17	0
18	289.66	68.42	15	76	530	1	0	1	0	2	0
19	881.24	69.17	15	67	1,090	0	0	0	0	1	0
20	490.88	68.92	16	59	1,050	1	0	0	0	8	0
21	567.79	68.75	11	70	913	0	0	1	1	15	0
22	665.99	70.92	22	57	828	1	1	0	0	20	0
23	621.45	69.67	16	59	786	0	0	1	0	18	0
24	608.8	70.08	19	58	821	1	0	0	0	3	0
25	473.64	70.42	19	44	538	0	0	1	0	19	0
26	697.14	71.08	20	57	1,130	0	0	1	0	21	0
27	207.51	67.25	13	63	745	0	0	0	0	8	1
28	288.48	67.17	9	48	821	0	0	1	0	7	1
29	284.88	67.83	12	63	886	0	0	0	1	11	1
30	280.36	67.83	12	71	886	1	0	0	1	11	1
31	217.38	67.25	13	72	745	1	0	0	0	8	1
32	270.71	67.83	7	80	886	1	0	0	1	11	1

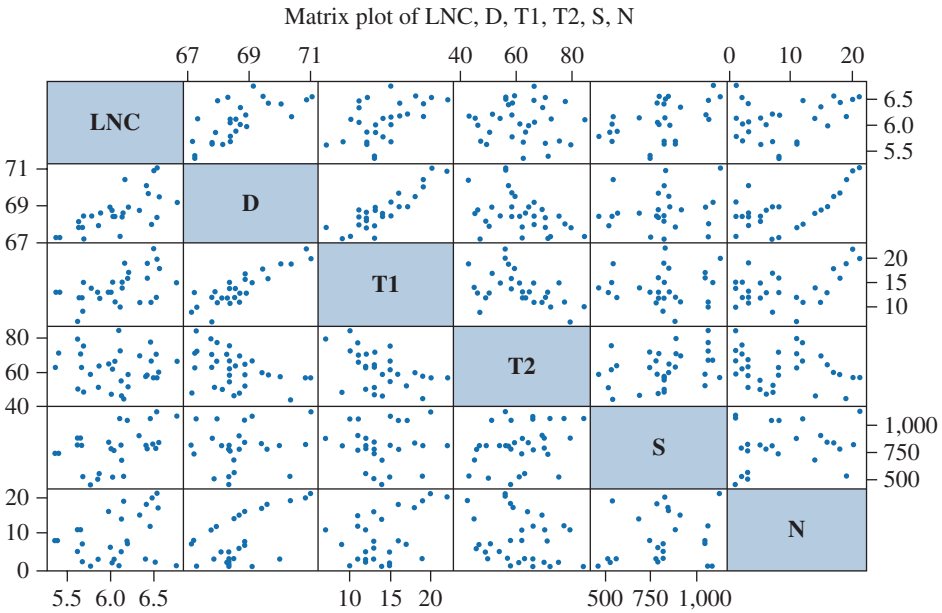
Source: *Cox and Snell (1981)*

The columns are identified according to the following notation:

C	Cost in dollars $\times 10^{-6}$, adjusted to 1976 base
D	Date construction permit issues (year. proportion of year)
T1	Time between application for and issue of permit
T2	Time between issue of operating license and construction permit
S	Power plant net capacity (MWe)
PR	Prior existence of an LWR on same site (=1)
NE	Plant constructed in northeast region of USA (=1)
CT	Use of cooling tower (=1)
BW	Nuclear steam supply system manufactured by Babcock–Wilcox (=1)
N	Cumulative number of power plants constructed by each architect–engineer
PT	Partial turnkey plant (=1)

Analyzing the Data

A preliminary analysis of the data and economic theory indicates that the variation in cost should increase with the value of the cost variable. This theory along with the data plots suggests that the log-transformation of cost (LNC = log(C)) yields a response variable that is more likely to satisfy the model conditions required for a regression analysis. A scatterplot matrix is given here.



From the plot, there appears to be a strong correlation between several of the explanatory variables. In particular, D and T1 appear to have a strong positive relationship and T1 and T2 appear to have a negative relationship. Because of the concern about the impact of collinearity on the fitted regression line, the correlation between the explanatory variables is given here. Note that the correlations are not computed with the variables PR, NE, CT, BW, and PT. all of these variables are indicator variables and their correlation with the other variables would not be meaningful.

```
Correlations: D, T1, T2, S, N

          D      T1      T2      S
T1    0.858
T2   -0.404  -0.474
S     0.020  -0.094   0.313
N     0.549   0.400  -0.228  0.193
```

From the above matrix, the only pair of variables that would indicate a potential problem is (T1, D), which has a correlation of .858. This value is just below our threshold value of .90, and, hence, both variables will be kept in the model. The above matrix does not detect correlations between various linear combinations of the variables. The following SAS output for the model of LNC regressed on the 10 explanatory variables includes values for VIF, studentized residuals, and Cook's *D*.

Dependent Variable: LNC

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	3.82363	0.38236	13.28	<.0001
Error	21	0.60443	0.02878		
Corrected Total	31	4.42806			

Root MSE	0.16965	R-Square	0.8635
Dependent Mean	6.06718	Adj R-Sq	0.7985
Coeff Var	2.79626		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-10.63398	5.71026	-1.86	0.0766	0
D	1	0.22760	0.08656	2.63	0.0157	8.31830
T1	1	0.00525	0.02230	0.24	0.8161	6.08159
T2	1	0.00561	0.00460	1.22	0.2360	2.45712
S	1	0.00088369	0.00018115	4.88	<.0001	1.26727
PR	1	-0.10813	0.08351	-1.29	0.2094	1.66568
NE	1	0.25949	0.07925	3.27	0.0036	1.30924
CT	1	0.11554	0.07027	1.64	0.1150	1.32422
BW	1	0.03680	0.10627	0.35	0.7326	1.91292
N	1	-0.01203	0.00783	-1.54	0.1394	2.64429
PT	1	-0.22197	0.13042	-1.70	0.1035	2.88092

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	6.1313	6.0046	0.0918	0.1268	0.143	0.889	| |* |	0.030
2	6.1159	6.1968	0.0870	-0.0809	0.146	-0.556	| *| |	0.010
3	6.0941	6.1560	0.0968	-0.0619	0.139	-0.444	| | |	0.009
4	6.4805	6.4481	0.0885	0.0324	0.145	0.224	| | |	0.002
5	6.4649	6.4017	0.1035	0.0633	0.134	0.471	| | |	0.012
6	5.8447	5.9721	0.0976	-0.1274	0.139	-0.918	| *| |	0.038
7	5.6072	5.8912	0.0794	-0.2840	0.150	-1.894	| ***| |	0.091
8	5.7596	5.7346	0.0822	0.0250	0.148	0.168	| | |	0.001
9	6.1249	5.8837	0.0946	0.2412	0.141	1.713	| |***	0.120
10	6.5370	6.4667	0.1278	0.0702	0.112	0.629	| |* |	0.047
11	5.8597	5.8555	0.0919	0.004216	0.143	0.0296	| | |	0.000
12	5.9979	6.2308	0.0940	-0.2329	0.141	-1.649	| ***| |	0.110
13	6.0215	5.9247	0.0832	0.0968	0.148	0.654	| |* |	0.012
14	6.2057	6.2768	0.0912	-0.0711	0.143	-0.497	| | |	0.009
15	5.9773	6.0805	0.1075	-0.1032	0.131	-0.786	| *| |	0.038
16	6.0481	6.0233	0.0872	0.0248	0.146	0.171	| | |	0.001
17	6.5685	6.4170	0.0979	0.1515	0.139	1.093	| |** |	0.054
18	5.6687	5.8950	0.1019	-0.2263	0.136	-1.669	| ***| |	0.143
19	6.7813	6.5148	0.1047	0.2665	0.134	1.996	| |***	0.223
20	6.1962	6.1906	0.0871	0.005567	0.146	0.0382	| | |	0.000
21	6.3418	6.2426	0.0981	0.0992	0.138	0.716	| |* |	0,023
22	6.5013	6.5851	0.1098	-0.0839	0.129	-0.649	| *| |	0.028
23	6.4321	6.2315	0.0812	0.2006	0.149	1.347	| |** |	0.049
24	6.4115	6.3226	0.1030	0.0889	0.135	0.660	| |* |	0.023
25	6.1604	6.1026	0.1053	0.0578	0.133	0.435	| | |	0.011
26	6.5470	6.8301	0.1163	-0.2831	0.124	-2.292	| ****| |	0.423
27	5.3352	5.4338	0.1055	-0.0987	0.133	-0.743	| *| |	0.032
28	5.6646	5.5053	0.1282	0.1594	0.111	1.435	| |** |	0.249
29	5.6521	5.6859	0.0960	-0.0338	0.140	-0.242	| | |	0.003
30	5.6361	5.6226	0.0923	0.0134	0.142	0.0944	| | |	0.000
31	5.3816	5.3762	0.1000	0.005483	0.137	0.0400	| | |	0.000
32	5.6010	5.6468	0.1259	-0.0458	0.114	-0.403	| | |	0.018

The values of VIF range from 1.3 to 8.3. Thus, no value is above 10, the value that would indicate a potential collinearity problem. Based on the scatterplot matrix, the values of the correlations, and the values of VIF, there does not appear to be any indication of collinearity. From the previous output, there is an indication of an outlier. The observation associated with plant 26 has a relatively large standardized residual, −2.292. However, the value of Cook's D is just .423, which would indicate that this observation does not have undue influence on the overall regression model.

The following output contains the results of fitting all possible regressions. Only the best (in terms of R^2) four models of each size, k, are displayed. There are substantial differences among the fits of the models with $k = 1, 2, 3$, and 4 variables. The maximum R^2_{adj} were .436, .631, .733, and .781 for $k = 1, 2, 3$, and 4, respectively. For the models with $k \geq 5$ variables in the model, the difference in maximum R^2_{adj} is much smaller, ranging from .798 for $k = 5$ to .815 for $k = 8$. For $k = 5$, the variables D, S, NE, CT, and PT yielded a model with $R^2_{adj} = .798$ and $s^2 = .0289$. For $k = 6$, the variables D, S, NE, CT, N, and PT yielded a model with $R^2_{adj} = .807$ and $s^2 = .0276$. The two models are not very different with respect to these two measures. For the models with more than seven variables, there is very little increase in R^2_{adj} or decrease in s^2. Thus, in terms of fit, the five-variable model with variables D, S, NE, CT, and PT provides nearly as good a fit as any of the models with six or more variables. An examination of the Mallow C_p values yields the following conclusions. The best five-variable model, $k = 5$, has $C_p = 6.06 \approx k + 1$. For models with $k < 5$, the C_p value associated with the best model of each size is larger than the desired value of $k + 1$. For example, with $k = 4$, $C_p = 7.30 > 5 = k + 1$. For models with $k > 5$, the C_p value associated with the best model of each size is smaller than the desired value of $k + 1$. For example, the best six-variable model, $k = 6$, has $C_p = 5.97$, which is less than $k + 1 = 7$.

```
                          Dependent Variable: LNC

                          R-Square Selection Method

                     Number of Observations Read          32
                     Number of Observations Used          32


Number in                 Adjusted
  Model     R-Square     R-Square        C(p)         MSE    Variables in Model

      1       0.4545       0.4364     55.9169     0.08051    PT
      1       0.3958       0.3756     64.9572     0.08918    D
      1       0.2064       0.1799     94.0922     0.11714    T1
      1       0.1963       0.1695     95.6423     0.11862    S
-------------------------------------------------------------------------------
      2       0.6552       0.6314     27.0471     0.05265    S PT
      2       0.5814       0.5525     38.4031     0.06392    D S
      2       0.5656       0.5357     40.8243     0.06632    D PT
      2       0.5530       0.5222     42.7701     0.06825    N PT
-------------------------------------------------------------------------------
      3       0.7585       0.7326     13.1611     0.03820    D S PT
      3       0.7167       0.6864     19.5836     0.04480    T1 S PT
      3       0.7088       0.6776     20.7990     0.04605    S N PT
      3       0.6989       0.6667     22.3210     0.04762    S NE PT
-------------------------------------------------------------------------------
      4       0.8096       0.7814      7.2969     0.03123    D S NE PT
      4       0.7821       0.7498     11.5221     0.03574    D S CT PT
      4       0.7640       0.7291     14.3043     0.03870    D T2 S PT
      4       0.7633       0.7283     14.4112     0.03882    T1 S NE PT
-------------------------------------------------------------------------------
```

5	0.8306	0.7980	6.0598	0.02885	D S NE CT PT
5	0.8216	0.7873	7.4447	0.03038	D T2 S NE PT
5	0.8177	0.7827	8.0448	0.03105	D S NE CT N
5	0.8150	0.7794	8.4660	0.03151	D S NE N PT
6	0.8442	0.8068	5.9732	0.02760	D S NE CT N PT
6	0.8376	0.7986	6.9822	0.02876	D T2 S NE CT PT
6	0.8368	0.7676	7.1098	0.02891	D T2 S PR NE PT
6	0.8335	0.7936	7.6115	0.02949	D S NE CT BW PT
7	0.8502	0.8065	7.0461	0.02764	D S NE CT BW N PT
7	0.8497	0.8059	7.1232	0.02773	D T2 S NE CT N PT
7	0.8483	0.8040	7.3459	0.02800	D S PR NE CT N PT
7	0.8472	0.8026	7.5058	0.02819	D T2 S PR NE CT PT
8	0.8627	0.8149	7.1296	0.02644	D T2 S PR NE CT N PT
8	0.8538	0.8029	8.4922	0.02815	D S PR NE CT BW N PT
8	0.8526	0.8013	8.6813	0.02838	D T2 S NE CT BW N PT
8	0.8506	0.7987	8.9809	0.02876	D T1 T2 S NE CT N PT
9	0.8631	0.8072	9.0555	0.02755	D T2 S PR NE CT BW N PT
9	0.8627	0.8066	9.1199	0.02763	D T1 T2 S PR NE CT N PT
9	0.8538	0.7940	10.4884	0.02942	D T1 S PR NE CT BW N PT
9	0.8526	0.7923	10.6766	0.02967	D T1 T2 S NE CT BW N PT
10	0.8635	0.7985	11.0000	0.02878	D T1 T2 S PR NE CT BW N PT

Based on the analysis given above, the model $\text{LNC} = \beta_0 + \beta_1\text{D} + \beta_2\text{S} + \beta_3\text{NE} + \beta_4\text{CT} + \beta_5\text{PT} + \varepsilon$ was fit to the data, yielding the following plots and summary information.

Dependent Variable: LNC

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3.67800	0.73560	25.50	<.0001
Error	26	0.75007	0.02885		
Corrected Total	31	4.42806			

Root MSE	0.16985	R-Square	0.8306
Dependent Mean	6.06718	Adj R-Sq	0.7980
Coeff Var	2.79948		

Parameter Estimates

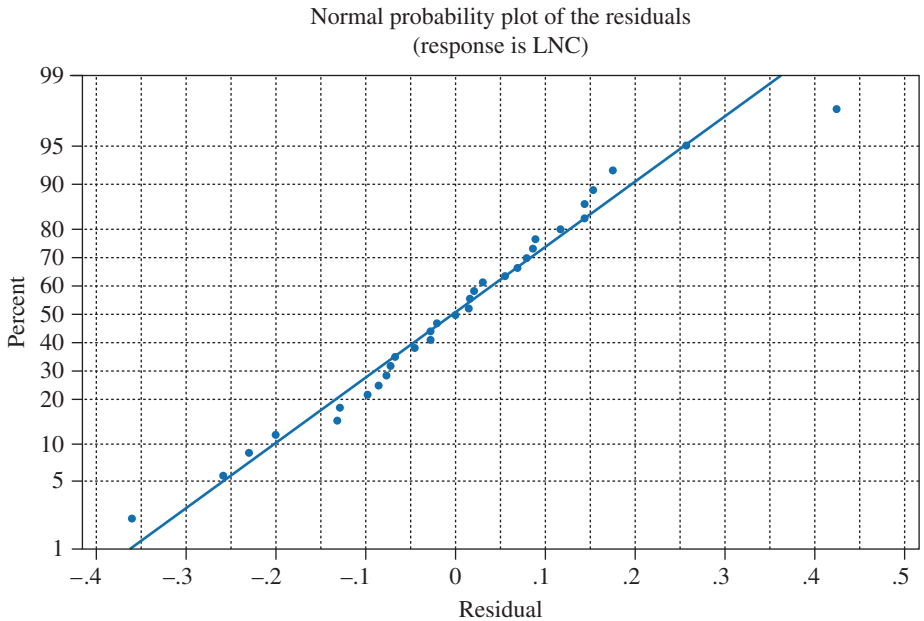
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-5.40584	2.45673	-2.20	0.0369	0
D	1	0.15640	0.03560	4.39	0.0002	1.40405
S	1	0.00086741	0.00016128	5.38	<.0001	1.00220
NE	1	0.19735	0.07233	2.73	0.0113	1.08796
CT	1	0.11542	0.06423	1.80	0.0839	1.10370
PT	1	-0.34777	0.09648	-3.60	0.0013	1.57282

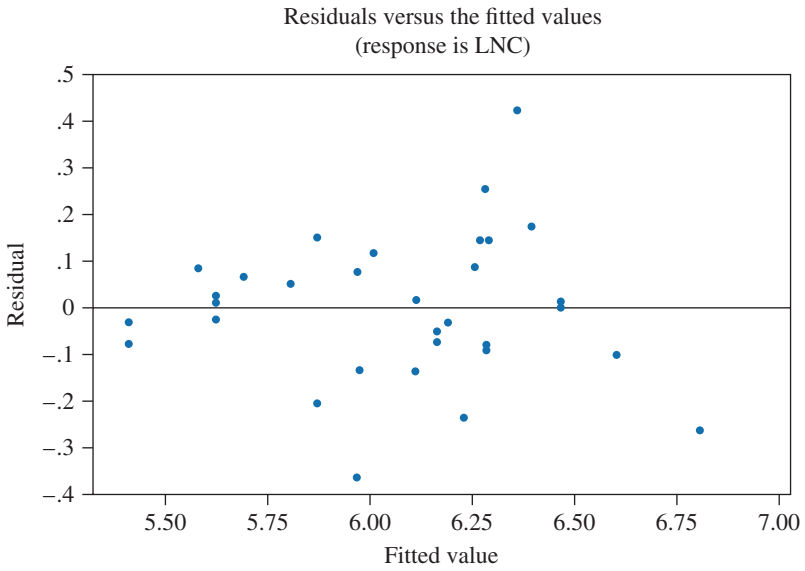
Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	6.1313	6.1133	0.0719	0.0181	0.154	0.117	| | |	0.001
2	6.1159	6.1637	0.0813	-0.0479	0.149	-0.321	| | |	0.005
3	6.0941	6.1637	0.0813	-0.0697	0.149	-0.467	| | |	0.011
4	6.4805	6.4659	0.0807	0.0147	0.149	0.0981	| | |	0.000
5	6.4649	6.4659	0.0807	-0.000926	0.149	-0.0062	| | |	0.000
6	5.8447	5.9754	0.0875	-0.1307	0.146	-0.898	| *| |	0.048

7	5.6072	5.9689	0.0573	-0.3617	0.160	-2.263	****	0.110
8	5.7596	5.6914	0.0786	0.0682	0.151	0.453		0.009
9	6.1249	6.0080	0.0532	0.1169	0.161	0.725	*	0.010
10	6.5370	6.2807	0.0686	0.2563	0.155	1.649	***	0.088
11	5.8597	5.8058	0.0659	0.0540	0.157	0.345		0.004
12	5.9979	6.2292	0.0682	-0.2313	0.156	-1.487	**	0.071
13	6.0215	5.8701	0.0724	0.1513	0.154	0.985	*	0.036
14	6.2057	6.2840	0.0615	-0.0783	0.158	-0.494		0.006
15	5.9773	6.1105	0.0492	-0.1332	0.163	-0.820	*	0.010
16	6.0481	5.9698	0.0536	0.0783	0.161	0.486		0.004
17	6.5685	6.3942	0.0709	0.1742	0.154	1.129	**	0.045
18	5.6687	5.8701	0.0724	-0.2014	0.154	-1.311	**	0.064
19	6.7813	6.3578	0.0656	0.4236	0.157	2.703	*****	0.213
20	6.1962	6.2840	0.0615	-0.0878	0.158	-0.554	*	0.008
21	6.3418	6.2540	0.0548	0.0878	0.161	0.546	*	0.006
22	6.5013	6.6016	0.0982	-0.1003	0.139	-0.724	*	0.044
23	6.4321	6.2877	0.0647	0.1444	0.157	0.919	*	0.024
24	6.4115	6.2668	0.0619	0.1447	0.158	0.915	*	0.021
25	6.1604	6.1899	0.0959	-0.0294	0.140	-0.210		0.003
26	6.5470	6.8066	0.1107	-0.2596	0.129	-2.016	****	0.500
27	5.3352	5.4105	0.0723	-0.0753	0.154	-0.490		0.009
28	5.6646	5.5793	0.0870	0.0853	0.146	0.585	*	0.020
29	5.6521	5.6235	0.0713	0.0286	0.154	0.185		0.001
30	5.6361	5.6235	0.0713	0.0126	0.154	0.0817		0.000
31	5.3816	5.4105	0.0723	-0.0288	0.154	-0.187		0.001
32	5.6010	5.6235	0.0713	-0.0224	0.154	-0.145		0.001

There are three plants that have somewhat large studentized residuals: plants 7, 19, and 26. However, Cook's D for the three plants is .110, .213, and .500. Therefore, the observations from these three plants do not have a large influence on the overall fit of the model. An assessment of the residuals from this model does not indicate the need for any higher-order or interaction terms in the five variables. The normal probability plot and a plot of residuals versus $\hat{y}$ are given here.





From the plots, there is no indication of a violation of the normality condition. There appears to be somewhat of an increase in the variance of the residuals for increasing values of the fitted values. However, the Breusch–Pagan test has a value of 5.61, which has a p-value of .23 in testing the null hypothesis of homogeneity of the variance. Thus, the constant variance condition does not appear to be violated. There is not apparent spatial or temporal ordering in the data, so it is not appropriate to test for serial correlation. Finally, the least-squares model computed from the data is

$$\hat{y} = -5.40584 + .15640\text{D} + .00086741\text{S} + .19735\text{NE} + .11542\text{CT} - .34777\text{PT}$$

Predicted construction costs can be computed from this equation, provided the values of D, S, NE, CT, and PT for the proposed plant fall within the space of these variables for the 32 plants used in the study. A more crucial conclusion from this study is the identification of those explanatory variables that most closely relate to construction costs. These variables can be used in planning the costs of constructing future plants.

13.6 Summary and Key Formulas

This key chapter presents some of the practical problems associated with multiple regression problems. Step 1 of the process is to decide on the dependent variable and a set of candidate independent variables for inclusion in the model. We discussed the invaluable nature of information from an expert in the subject matter field and the utility of some of the best subset regression techniques for choosing which variables to include in the model.

Step 2 involves the actual polynomial form of the particular multiple regression equation. In particular, attention should be paid to the lack of fit of a proposed model to the data collected on the dependent and independent variables of interest. A formal test for lack of fit of a polynomial model is possible where there are repetitions of observations at one or more settings of the independent variables. Lack of fit can also be examined using residual plots.

Following steps 1 and 2 as we've discussed them can sometimes be a problem depending on the data that are available. For example, if data are available on many variables at the time that the multiple regression model is being formulated, then consultation with experts and application of one (or more) of the best subset regression techniques can be useful in culling the list of potential independent variables (step 1). The regression model is then modified in step 2 based on the discussions and analyses of step 1. Sometimes, however, data are not available on many possible independent variables. For these situations, step 1 consists of discussions with experts to determine which variables may be important predictors; data are then gathered on these variables. After the data are obtained on these candidate independent variables, the subset regression techniques and the model formulation techniques of step 2 can be applied to refine the model.

The final step of the multiple regression problem is to check the underlying assumptions of multiple regression: zero expectation, constant variance, normality, and independence. Although some formal tests were presented, violation of the assumption is checked best by closely examining the data using scatterplots, various residual plots, and normal probability plots. The more experience one gains in examining and interpreting data with these plots, the better will be the resulting regression equations.

Key Formulas

1. C_P statistic

$$C_P = \frac{\text{SS(Residual)}_p}{s_\varepsilon^2} - (n - 2p)$$

2. AIC statistic

$$\text{AIC}_k = n\log_e(\text{SS(Residual)}/n) + 2k$$

3. BIC statistic

$$\text{BIC}_k = n\log_e(\text{SS(Residual)}/n) + k\log_e(n)$$

4. Backward elimination

$$F_j = \frac{\text{SSR}_j - \text{SSR}}{\text{MS(Residual)}}, \quad j = 1, 2, \ldots$$

5. Durbin–Watson statistic

$$d = \frac{\sum_{t=1}^{n-1}(e_{t+1} - e_t)^2}{\sum_t e_t^2}$$

13.7 Exercises

13.2 Selecting the Variables (Step 1)

Edu.

13.1 A recent lawsuit addressed the issue of whether student-athletes should receive a stipend above the costs of tuition and room and board to compensate them for the enormous amounts of money generated by university athletic departments using the student-athletes' images in video games and other such products. The NCAA wants to examine the economic feasibility of this added expense of intercollegiate athletics for the universities under its jurisdiction.

What explanatory variables would be useful in predicting whether universities would be able to support this added expense? It may be useful to have several dummy variables in your model.

H.R. **13.2** A large computer software firm wants to evaluate its employees' satisfaction with the immediate supervisors. A survey is to be designed to make this assessment. The company wants to relate the overall rating of the performance of the supervisor, as ascertained from a survey question, to explanatory variables that can be obtained from questions in the survey and the employees' personnel files. What questions would you include in the survey? What personal information about the employees would be pertinent? Propose a model that would use the information obtained from and about each employee to predict the employee's satisfaction with their supervisor.

Soc. **13.3** A sociologist is studying what factors may affect whether college students would support new laws that would make it a crime for students to purchase papers from the Internet and then turn in the papers as their own work. A random sample of 45 students at a large state university is interviewed and asked to provide a measure of their strength of support for criminalizing the purchase of term papers. A CRIME score from 0 to 25 is obtained from each student, with 0 being totally opposed to criminal penalties and 25 being totally in favor of criminal penalties. Information on the following explanatory variables was also obtained from each student: age of student (A), number of years of college (C), income of parents (I) (in $1,000), and gender (G) (0 = female).

The data are shown here:

Stu	CRIME	A	C	I	G	Stu	CRIME	A	C	I	G
1	2	16	2	83	1	24	0	32	4	72	1
2	0	18	2	92	1	25	3	32	4	75	1
3	3	18	2	95	1	26	0	31	4	77	0
4	9	18	2	81	0	27	8	30	4	66	1
5	6	19	2	85	1	28	11	29	4	55	0
6	6	19	2	90	1	29	13	29	4	52	0
7	7	20	2	98	1	30	15	28	4	50	0
8	9	19	2	96	0	31	17	27	4	49	0
9	13	18	2	73	0	32	18	26	4	48	0
10	12	19	2	76	0	33	20	25	4	45	0
11	9	19	2	79	1	34	16	24	3	53	0
12	12	20	2	75	0	35	18	23	3	46	0
13	12	21	2	80	0	36	16	23	3	48	1
14	11	20	2	72	0	37	15	22	3	58	0
15	11	24	3	74	0	38	21	22	3	44	0
16	12	25	3	75	0	39	19	22	3	48	0
17	9	25	3	75	1	40	17	21	3	49	1
18	9	27	4	76	1	41	14	21	2	55	1
19	11	28	4	72	0	42	15	20	2	53	0
20	5	38	4	79	0	43	19	19	2	47	0
21	0	29	4	83	1	44	18	21	3	44	0
22	6	30	4	75	1	45	10	21	2	73	1
23	2	31	4	79	0						

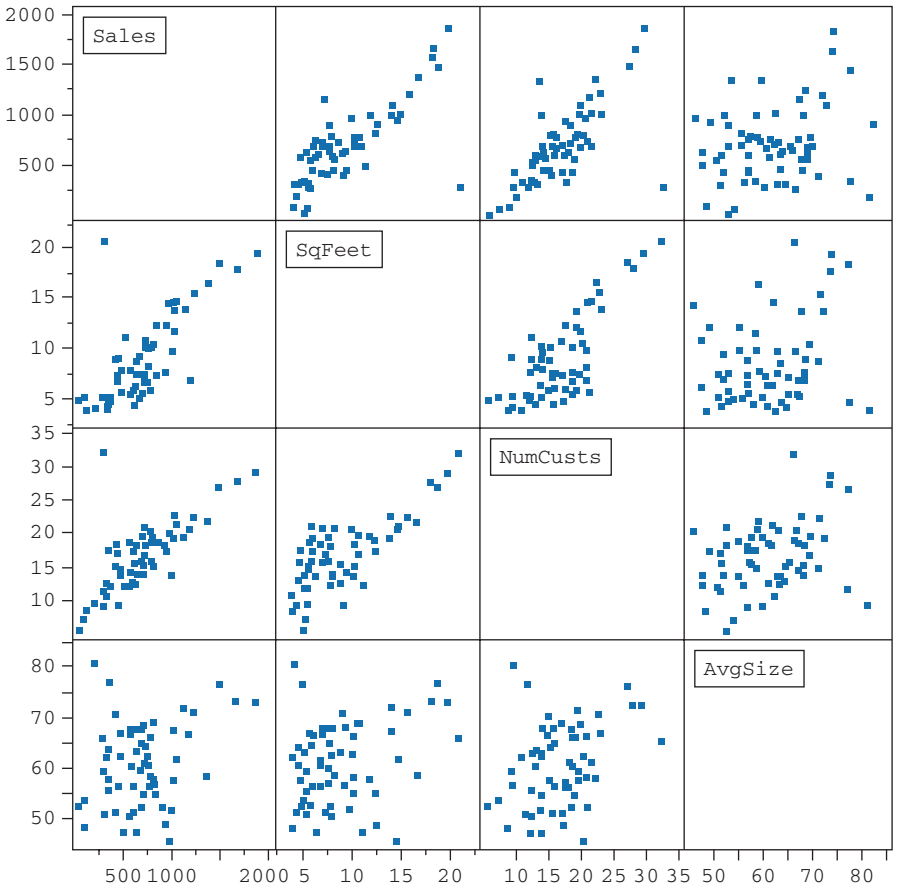
a. Are there any collinearity problems based on the above data?

b. Use the output from a best subset regression software program to determine which explanatory variables should be included in the model.

c. What other explanatory variables may have been related to the response variable CRIME?

13.4 Refer to Exercise 13.3. Use the output from a stepwise regression software program to determine which explanatory variables should be included in the model. Compare the results of your conclusions from the stepwise program to your results from the best subset program.

Bus. **13.5** A supermarket chain staged a promotion for organic vegetables. The actual sales (Sales) of organic vegetables for the weekend of the promotion were obtained from scanner data at the checkout. Three explanatory variables under consideration for modeling SALES were the size of the store (SqFeet) (in thousands of square feet), the number of customers processed in the store (NumCusts) (in hundreds), and the average size of purchase (AvgSize), which was also obtained from the scanner data. A scatterplot matrix is shown here.



a. Is there any evidence of collinearity in the scatterplots?
b. Does the scatterplot matrix reveal any other problems associated with the data?
c. What other diagnostics of collinearity would you suggest for this problem?

Engin. **13.6** The basic process of making paper has not changed in more than 2,000 years. It involves two stages: the breaking up of raw material in water to form a suspension of individual fibers and the formation of felted sheets by spreading this suspension on a suitable porous surface, through which excess water can drain. Most paper is made from wood pulp that has been bleached with chlorine. This bleaching takes place for two reasons: to remove the last traces of a material called lignin from the raw pulp in order to make the paper stronger and to create a brilliant white writing surface. Chlorine is an ideal chemical for these tasks, but unfortunately its use in paper mills also results in a wide variety of toxic substances being released into the environment. Studies have been conducted to determine which factors in the paper process are most highly correlated with

the brightness of finished paper. The article ***"Advantages of CE-HDP Bleaching for High Brightness Kraft Pulp production" [Tappi (1964) 47:170A–175A]*** contains the following data on these variables: y = brightness of finished paper, x_1 = hydrogen peroxide (% by weight), x_2 = sodium hydroxide (% by weight), x_3 = silicate (% by weight), and x_4 = process temperature (in °F). There were 31 runs in the study.

Run	x_1	x_2	x_3	x_4	y	Run	x_1	x_2	x_3	x_4	y
1	.2	.2	1.5	145	83.9	17	.1	.3	2.5	160	82.9
2	.4	.2	1.5	145	84.9	18	.5	.3	2.5	160	85.5
3	.2	.4	1.5	145	83.4	19	.3	.1	2.5	160	85.2
4	.4	.4	3.5	145	84.2	20	.3	.5	2.5	160	84.5
5	.2	.2	3.5	145	83.8	21	.3	.3	2.5	160	84.7
6	.4	.2	3.5	145	84.7	22	.3	.3	2.5	160	85.0
7	.2	.4	3.5	145	84.0	23	.3	.3	2.5	160	84.9
8	.4	.4	1.5	175	84.8	24	.3	.3	2.5	160	84.0
9	.2	.2	1.5	175	84.5	25	.3	.3	2.5	160	84.5
10	.4	.2	1.5	175	86.0	26	.3	.3	2.5	160	84.7
11	.2	.4	1.5	175	82.6	27	.3	.3	2.5	160	84.6
12	.4	.4	3.5	175	85.1	28	.3	.3	2.5	160	84.9
13	.2	.2	3.5	175	84.5	29	.3	.3	2.5	160	84.9
14	.4	.2	3.5	175	86.0	30	.3	.3	2.5	160	84.5
15	.2	.4	3.5	175	84.0	31	.3	.3	2.5	160	84.6
16	.4	.4	3.5	175	85.4						

a. Use scatterplots and VIF to determine if there is evidence of collinearity in the explanatory variables.
b. This was a designed experiment with nonrandom explanatory variables. Was it really necessary to investigate collinearity in this type of study?
c. Use a variable selection procedure with minimum BIC as the criterion to formulate a model.
d. Use a variable selection procedure with maximum R^2_{adj} as the criterion to formulate a model.
e. Compare the results of parts (c) and (d).

13.7 Refer to Exercise 13.6. Include the square of each of the explanatory variables and all cross-product terms in your model selection procedure.

a. Use a variable selection procedure with maximum BIC R^2_{adj} as the criterion to formulate a model.
b. Use a variable selection procedure C_p as the criterion to formulate a model.
c. Use a variable selection procedure with minimum BIC statistic as the criterion to formulate a model.
d. Compare the included terms from the models formulated with the three criteria in parts (a)–(c).

13.3 Formulating the Model (Step 2)

Ag.

13.8 The cotton aphid is pale to dark green in cool seasons and yellow in hot, dry summers. Generally distributed throughout temperate, subtropic, and tropic zones, the cotton aphid occurs in all cotton-producing areas of the world. These insects congregate on lower leaf surfaces and on terminal buds, extracting plant sap. If weather is cool during the spring, populations of natural enemies will be slow in building up, and heavy infestations of aphids may result. When this occurs, leaves begin to curl and pucker; seedling plants become stunted and may die. Most aphid damage is of this type. If honeydew resulting from late-season aphid infestations falls onto open cotton, it can act as a growing medium for sooty mold. Cotton stained by this black fungus is reduced

in quality and brings a low price for the grower. Entomologists studied the aphids to determine weather conditions that may result in increased aphid density on cotton plants. The following data were reported in **Statistics and Data Analysis *(Peck, Olson, and Devore, 2005)*** and come from an extensive study as reported in the article ***"Estimation of the Economic Threshold of Infestation for Cotton Aphid"* [Mesopotamia Journal of Agriculture *(1982): 10, 71–75*].** In the following table,

y = infestation rate (aphids/100 leaves)
x_1 = mean temperature (°C)
x_2 = mean relative humidity

Field	y	x_1	x_2	Field	y	x_1	x_2
1	61	21.0	57.0	18	25	33.5	18.5
2	77	24.8	48.0	19	67	33.0	24.5
3	87	28.3	41.5	20	40	34.5	16.0
4	93	26.0	56.0	21	6	34.3	6.0
5	98	27.5	58.0	22	21	34.3	26.0
6	100	27.1	31.0	23	18	33.0	21.0
7	104	26.8	36.5	24	23	26.5	26.0
8	118	29.0	41.0	25	42	32.0	28.0
9	102	28.3	40.0	26	56	27.3	24.5
10	74	34.0	25.0	27	60	27.8	39.0
11	63	30.5	34.0	28	59	25.8	29.0
12	43	28.3	13.0	29	82	25.0	41.0
13	27	30.8	37.0	30	89	18.5	53.5
14	19	31.0	19.0	31	77	26.0	51.0
15	14	33.6	20.0	32	102	19.0	48.0
16	23	31.8	17.0	33	108	18.0	70.0
17	30	31.3	21.0	34	97	16.3	79.5

a. Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ to the aphid data.
b. Use residual plots, tests of hypotheses, and other diagnostic statistics to identify possible additional terms to add to the model fit in part (a).

13.9 Refer to Exercise 13.8.
a. Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$ to the aphid data.
b. Compare the fit of the linear model from Exercise 13.8 to the fully quadratic model fit in part (a) of this exercise.
c. Use residual plots, tests of hypotheses, and other diagnostic statistics to identify possible additional terms to add to the model fit in part (a).

13.10 Refer to Exercise 13.9.
a. What is the incremental increase to R^2 for the model of Exercise 13.8 as opposed to the model considered in part (a) of Exercise 13.9?
b. Is this incremental increase statistically significant as measured by an F test at $\alpha = .05$?

13.11 Refer to Exercise 13.8.
a. Take as the response variable $ty = \log(y)$, the natural logarithm of the aphid count. Fit the model $ty = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ to the aphid data.
b. Compare the fit of the quadratic model from Exercise 13.9 to the linear model fit in part (a) of this exercise.
c. Can we validly compare the R^2_{adj} values from these two models? Justify your answer.

Bus. **13.12** A business analyst at a major real estate firm wants to build a regression model that will predict the price of a single-family home based on a number of explanatory variables. The real estate firm's data base has an enormous amount of information on selling prices of homes and potential variables that are related to that price. After a number of discussions with experienced realtors, you decide to model price of a home as a function of the following variables: size of home in terms of floor space, lot size, number of bathrooms, number of bedrooms, age of home, garage size, number of months home has been on the market, distance to nearest elementary school, type of neighborhood, traffic volume on street, and racial mixture of neighborhood.

a. Would you expect any of these variables to be highly correlated?
b. How would you determine if there was a high correlation?
c. What impact would a high correlation between independent variables have on the fitted regression?

13.13 Refer to Exercise 13.12. The analyst proposed the following variable to describe the type of roof on the home:

$$\text{RoofType} = \begin{cases} 3 \text{ if roof has asphalt shingles} \\ 2 \text{ if roof has metal shingles} \\ 1 \text{ if roof has cedar shingles} \\ 0 \text{ otherwise} \end{cases}$$

a. Discuss any problems with this variable.
b. How could type of roof be included in the model so as not to pose the problems associated with the above definition.

13.14 Refer to Exercise 13.12. The realtors informed the analyst that the relation between selling price and age of home could vary greatly depending on the type of roof. What terms would need to be included in the regression model in order to be able to evaluate whether the relation between selling price and age of home varies depending on the type of roof?

13.15 Refer to Exercise 13.12. The realtors suspect that the impact on selling price of increasing age of home is itself increasing. That is, there is little difference in the selling prices of homes with age = 1 year to 10 years, but a larger decrease for homes from 11 years to 20 years old, a very large decrease for homes from 21 to 30 years old, and so on.

a. What terms would be needed in the regression model to evaluate whether the realtors' suspicion is valid?
b. If the realtors' suspicion is true, what type of pattern would you expect to see in a plot of the residuals versus age of home from a regression model having just a first order term in age of home?

13.16 Refer to Exercise 13.12. After assembling the data set and fitting the regression model to the data, the analyst realizes that a number of the homes appear in the data set multiple times because the data set contains the selling price of homes over the past 20 years and a number of the homes had been sold multiple times. What types of problems would this cause in the regression model, and how could these problems be addressed?

Ag. **13.17** Hops originate from the flowers of Humulus lupulus and are used primarily as a flavoring and stability agent in beer. Hops have several characteristics that are very favorable to beer: Hops contribute a bitterness that balances the sweetness of the malt, hops can contribute aromas, and hops have an antibiotic effect that favors the activity of brewer's yeast over less desirable microorganisms. The bitterness level of a particular hop variety is measured in percent alpha acid by weight. The higher the percentage, the more bitter the hop in direct proportion. Alpha acids are now the accepted method in the brewing industry for assessing the quality of hops. The European Brewery Company carried out trials in six countries on four varieties of hops to determine if the mean temperature and mean duration of sunshine between the date of the flower coming into hop and the date of picking (the critical dates) have an impact on the alpha acid content of hops. The following data were reported by ***Smith*** in the article ***"The Influence of Temperature and Sunshine on the Alpha-Acid Content of Hops"* [Agricultural Meteorology *(1974) 13:375–382].** The variables in the following table are P (alpha acid %), T (mean temperature, °C), and S (mean sunshine, h/day), where the means are over the critical dates. There were four varieties of hops included in the study.

Variety of Hops															
Fuggle				Northern Brewer				Hallertau				Saaz			
Field	*P*	*T*	*S*	Field	*P*	*T*	*S*	Field	*P*	*T*	*S*	Field	*P*	*T*	*S*
1	7.2	16.7	4.4	1	12.1	16.8	4.4	1	5.5	16.5	4.4	1	6.8	16.7	4.4
2	5.8	17.4	5.8	2	10.7	17.0	6.2	2	5.3	17.1	5.8	2	4.9	18.5	7.5
3	5.7	17.1	5.9	3	10.6	17.9	5.9	3	4.7	18.4	7.0	3	4.7	18.1	7.8
4	5.5	18.9	6.2	4	10.2	18.0	7.7	4	4.6	17.4	5.8	4	4.6	17.1	5.7
5	5.2	17.7	6.6	5	9.6	18.0	6.9	5	4.5	18.3	7.5	5	4.1	18.7	7.1
6	5.1	18.4	6.9	6	9.1	21.3	6.1	6	4.4	18.6	7.5	6	3.9	17.9	5.9
7	4.8	16.8	6.9	7	8.8	18.5	7.2	7	4.0	19.3	6.7	7	3.8	19.1	7.1
8	4.8	18.2	6.2	8	8.8	19.1	6.5	8	3.8	19.2	6.5	8	3.5	21.4	5.9
9	4.8	20.7	8.4	9	8.1	19.9	8.5	9	3.2	21.4	6.1	9	3.4	19.0	7.6
10	4.7	21.3	6.2	10	8.0	19.1	6.6	10	3.3	20.6	8.7	10	3.1	17.7	7.1
11	4.3	21.2	7.4	11	7.6	21.1	7.3	11	3.0	19.8	8.5	11	3.0	20.9	7.8
12	3.7	17.3	6.9	12	6.4	17.4	6.9	12	2.9	21.2	7.9	12	2.7	19.0	8.8
13	3.2	18.5	8.6	13	5.8	19.2	8.4	13	2.8	17.3	6.9	13	2.5	20.1	8.5

a. Fit the model $P = \beta_0 + \beta_1 T + \beta_2 S + \varepsilon$ to the hops data with a separate equation for each variety.

b. Use residual plots, tests of hypotheses, and other diagnostic statistics to identify possible additional terms to add to the four models fit in part (a).

13.18 Refer to Exercise 13.17.

a. Using an indicator variable, fit a single model to the hops data for varieties Fuggle and Northern Brewer.

b. Using your results from part (a), obtain separate prediction equations for varieties Fuggle and Northern Brewer.

c. Interpret the values of the coefficients (βs) in the model.

d. Using your prediction equations in part (b), estimate the mean alpha acid percentage when the atmospheric conditions are a mean temperature of 19°C and a mean sunshine of 6.5. How different are the two estimates?

e. Place 95% confidence intervals on your estimates.

13.19 Refer to Exercise 13.17.

a. Using an indicator variable, fit a single model to the hops data for varieties Hallertau and Saaz.

b. Using your results from part (a), obtain separate prediction equations for varieties Hallertau and Saaz.

c. Interpret the values of the coefficients (βs) in the model.

d. Using your prediction equations in part (b), estimate the mean alpha acid percentage when the atmospheric conditions are a mean temperature of 19°C and a mean sunshine of 6.5. How different are the two estimates?

e. Place 95% confidence intervals on your estimates.

13.20 Refer to Exercise 13.17.

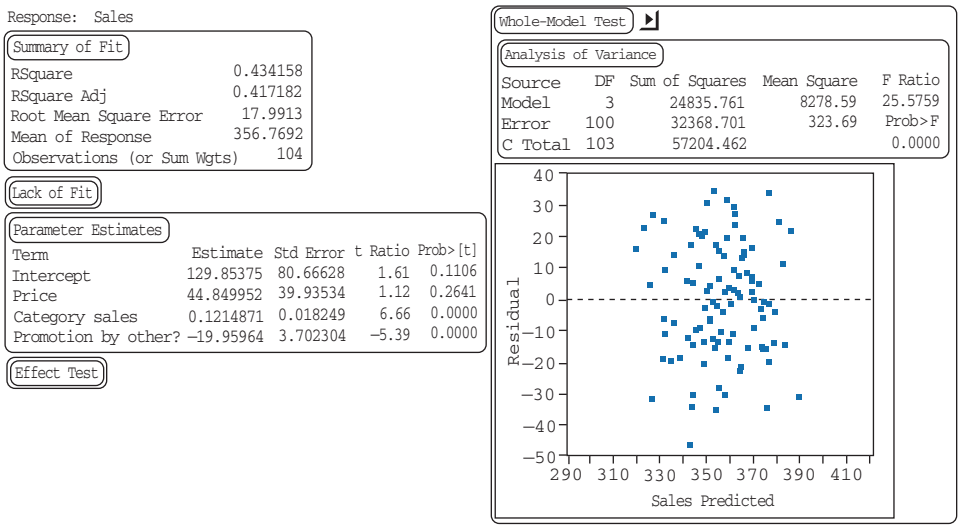
a. Using the model fit in part (a) of Exercise 13.18, is there significant evidence ($\alpha = .05$) that the mean sunshine partial slope coefficients are different?

b. Using the model fit in part (a) of Exercise 13.19, is there significant evidence ($\alpha = .05$) that the mean temperature partial slope coefficients are different?

c. Interpret the values of the coefficients (βs) in the model.

Bus. **13.21** A supermarket chain analyzed data on sales of a particular brand of snack cracker at 104 stores in the chain for a certain 1-week period. The analyst tried to predict sales based on the total sales of all brands in the snack cracker category, the price charged for the particular brand in question, and whether or not there was a promotion for a competing brand at a given store (promotion = 1 if there was such a promotion, 0 if not). (There were no promotions for the brand in question.) A portion of the JMP multiple regression output is shown in the figure.

a. Interpret the coefficient of the promotion variable.
b. Should a promotion by a competing product increase or decrease sales of the brand in question? According to the coefficient, does it?
c. Is the coefficient significantly different from 0 at usual α values?



Response: Sales

Summary of Fit

RSquare	0.434158
RSquare Adj	0.417182
Root Mean Square Error	17.9913
Mean of Response	356.7692
Observations (or Sum Wgts)	104

Lack of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	129.85375	80.66628	1.61	0.1106
Price	44.849952	39.93534	1.12	0.2641
Category sales	0.1214871	0.018249	6.66	0.0000
Promotion by other?	−19.95964	3.702304	−5.39	0.0000

Effect Test

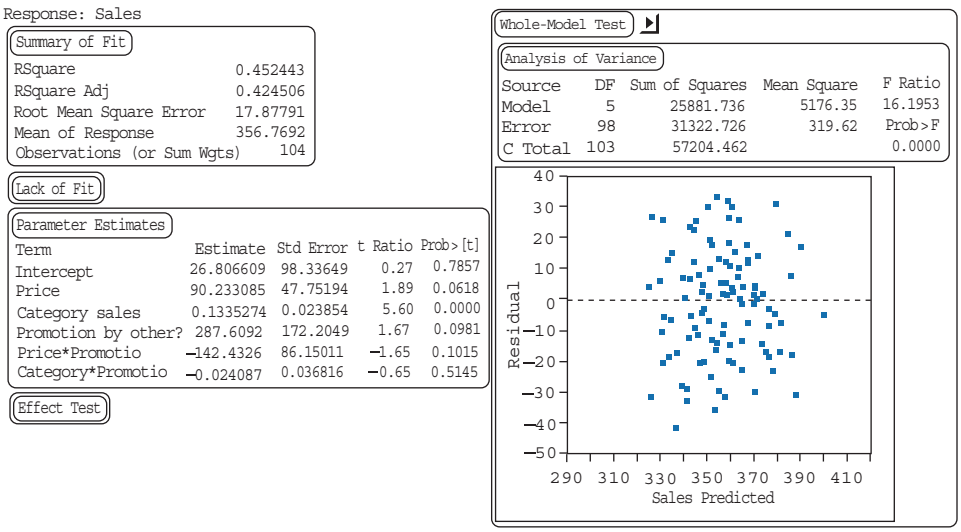
Whole-Model Test

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	24835.761	8278.59	25.5759
Error	100	32368.701	323.69	Prob>F
C Total	103	57204.462		0.0000

13.22 Refer to Exercise 13.21. How accurately can sales be predicted for a particular week, with 95% confidence?

Bus. **13.23** Refer to Exercise 13.21. An additional regression model for the snack cracker data is run, incorporating products of the promotion variable with price and with category sales. The output for this model is given in the figure. What effect do the product term coefficients have in predicting sales when there is a promotion by a competing brand? In particular, do these coefficients affect the intercept of the model or the slopes?



Response: Sales

Summary of Fit

RSquare	0.452443
RSquare Adj	0.424506
Root Mean Square Error	17.87791
Mean of Response	356.7692
Observations (or Sum Wgts)	104

Lack of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	26.806609	98.33649	0.27	0.7857
Price	90.233085	47.75194	1.89	0.0618
Category sales	0.1335274	0.023854	5.60	0.0000
Promotion by other?	287.6092	172.2049	1.67	0.0981
Price*Promotio	−142.4326	86.15011	−1.65	0.1015
Category*Promotio	−0.024087	0.036816	−0.65	0.5145

Effect Test

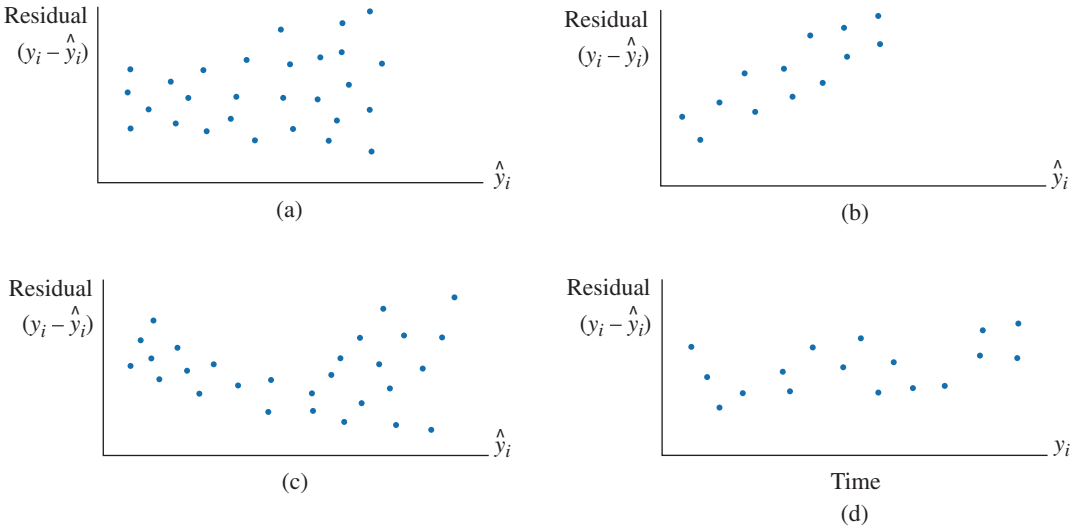
Whole-Model Test

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	25881.736	5176.35	16.1953
Error	98	31322.726	319.62	Prob>F
C Total	103	57204.462		0.0000

13.4 Checking Model Assumptions (Step 3)

13.24 Several different patterns of residuals are shown in the following plots. Indicate whether the plot suggests a problem, and, if so, indicate the potential problem and a possible solution.



Bus. **13.25** The book **Small Data Sets** *reports on an article by Kadiyala, "Testing for the Independence of Regression Disturbances" [Econometrica (1970) 38:97–117].* This article contains information on ice cream consumption over 30 4-week periods from March through July. The researchers were interested in determining what explanatory variables impacted the level of consumption. The variables considered in the study are

y, ice cream consumption, pints per capita
x_1, price of ice cream, $ per pint
x_2, weekly family income, $
x_3, mean temperature, °F

Period	y	x_1	x_2	x_3	Period	y	x_1	x_2	x_3
1	.386	.270	78	41	16	.381	.287	82	63
2	.374	.282	79	56	17	.470	.280	80	72
3	.393	.277	81	63	18	.443	.277	78	72
4	.425	.280	80	68	19	.386	.277	84	67
5	.406	.272	76	69	20	.342	.277	86	60
6	.344	.262	78	65	21	.319	.292	85	44
7	.327	.275	82	61	22	.307	.287	87	40
8	.288	.267	79	47	23	.284	.277	94	32
9	.269	.265	76	32	24	.326	.285	92	27
10	.256	.277	79	24	25	.309	.282	95	28
11	.286	.282	82	28	26	.359	.265	96	33
12	.298	.270	85	26	27	.376	.265	94	41
13	.329	.272	86	32	28	.416	.265	96	52
14	.318	.287	83	40	29	.437	.268	91	64
15	.381	.277	84	55	30	.548	.260	90	71

a. Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to the ice cream data. Is there evidence in the residual plots of serial correlation?

b. Perform a Durbin–Watson test for serial correlation. Does the test confirm your observations from the residual plots?

13.26 Refer to Exercise 13.25. Form first differences in the data and then regress the y differences on the x differences.

a. Is there evidence in the residual plots of serial correlation?
b. Perform a Durbin–Watson test for serial correlation. Does the test confirm your observations from the residual plots?

13.27 Refer to the crime data in Exercise 13.3. Obtain the residuals from the model you selected in Exercise 13.3.
a. Is there evidence in the residuals of a violation of the normality condition?
b. Is there evidence in the residual plots of a violation of the constant variance condition?
c. Perform a BP test for constant variance. Does the test agree with your observations in part (a)?
d. Determine the appropriate Box–Cox transformation for this data.

13.28 Refer to the papermaking data in Exercise 13.6. Obtain the residuals from the model you selected in Exercise 13.6.
a. Is there evidence in the residuals of a violation of the normality condition?
b. Is there evidence in the residual plots of a violation of the constant variance condition?
c. Perform a BP test for constant variance. Does the test agree with your observations in part (a)?
d. Determine the appropriate Box–Cox transformation for these data.

13.29 Refer to the aphid data in Exercise 13.8. Obtain the residuals from the model you selected in Exercise 13.9.
a. Is there evidence in the residuals of a violation of the normality condition?
b. Is there evidence in the residual plots of a violation of the constant variance condition?
c. Perform a BP test for constant variance. Does the test agree with your observations in part (a)?
d. Determine the appropriate Box–Cox transformation for these data.

13.30 Refer to the hops data in Exercise 13.17. Obtain the residuals from each of the four models you selected in Exercise 13.17.
a. Is there evidence in the residuals of a violation of the normality condition?
b. Is there evidence in the residual plots of a violation of the constant variance condition?
c. Perform a BP test for constant variance. Does the test agree with your observations in part (a)?
d. Determine the appropriate Box–Cox transformation for these data.

Soc. **13.31** A researcher in the social sciences examined the relationship between the rate (per 1,000) of nonviolent crimes y based on the rate of nonviolent crimes 5 years ago x_1 and the present unemployment rate x_2 for cities. Data from 20 different cities are shown here.

CITY	PRESENT RATE	RATE 5 YEARS AGO	PRESENT UNEMPLOYMENT RATE
1	13	14	5.1
2	8	10	2.7
3	14	16	4.0
4	10	10	3.4
5	12	16	3.1
6	11	12	4.3
7	7	8	3.8
8	6	7	3.2
9	10	12	3.2
10	16	20	4.1
11	16	14	5.9
12	9	10	4.0
13	11	10	4.1
14	18	20	5.0
15	9	13	3.1
16	10	6	6.3
17	15	10	5.7
18	14	14	5.2
19	17	16	4.9
20	6	8	3.0

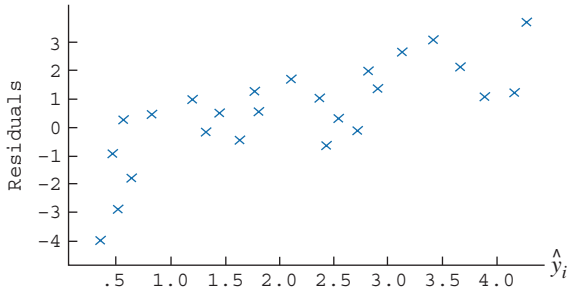
a. Determine the fit to the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

b. Examine the assumptions underlying the regression model. Discuss whether the assumptions appear to hold. If they don't, suggest possible remedies.

13.32 Refer to Exercise 13.31. Predict the present crime rate for a city having a crime rate of 9 (per 1,000) 5 years ago and an unemployment rate of 16%. Might there be a problem with this prediction? If so, why?

13.33 Estimates ($\hat{y}$s) and residuals from a securities firm's regression model for the prediction of earnings per share (per quarter) are shown here for 25 different high-technology companies. Is there any evidence that the assumptions have been violated? Are any additional tests or plots warranted?



Supplementary Exercises

Sci.

13.34 A construction science researcher is interested in evaluating the relationship between energy consumption by the homeowner and the difference between the internal and external temperatures. There were 30 homes used in the study. During an extended period of time, the average temperature difference (in °F) inside and outside the homes was recorded. The average energy consumption was also recorded for each home. The data are given here with y = energy consumption and x = mean temperature difference. Plot the data and suggest a polynomial model between y and x.

y	16	12	7	40	26	33	98	105	65	130	90	109	101	118	123
x	1	1	1	3	3	3	6	6	6	9	9	9	12	12	12
y	99	113	105	90	109	115	134	105	129	119	133	99	195	149	160
x	15	15	15	18	18	18	21	21	21	24	24	24	30	30	30

13.35 Refer to the data of Exercise 13.34.

a. Fit a cubic model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$.

b. Test for lack of fit of the model at the $\alpha = .05$ level.

c. Evaluate the normality and constant variance assumptions.

13.36 Refer to Exercise 13.34. As happens in many studies, not all the data are correctly collected. The researcher decides that errors are present in the information collected at several of the homes. After eliminating the questionable data values, the data appropriate for modeling are given here.

y	16	12	7	40	26	33	105	65	130	101	118	123
x	1	1	1	3	3	3	6	6	9	12	12	12
y	99	113	105	109	115	134	105	133	99	195	149	160
x	15	15	15	18	18	21	21	24	24	30	30	30

a. Fit a cubic model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$ to the reduced data set.

b. Compare the fit of the model in Exercise 13.35 to the fit of the model in part (a).

Med. **13.37** A pharmaceutical firm wanted to obtain information on the relationship between the dose level of a drug product and its potency. To do this, each of 15 test tubes was inoculated with a virus culture and incubated for 5 days at 30°C. Three test tubes were randomly assigned to each of the five different dose levels to be investigated (2, 4, 8, 16, and 32 mg). Each tube was injected with only one dose level, and the response of interest (a measure of the protective strength of the product against the virus culture) was obtained. The data are given here.

Dose Level	Response
2	5, 7, 3
4	10, 12, 14
8	15, 17, 18
16	20, 21, 19
32	23, 24, 29

a. Plot the data.
b. Fit both a linear and a quadratic model to these data.
c. Which model seems more appropriate?

Med. **13.38** Refer to Exercise 13.37. A logarithmic transformation of the dose levels will often result in a linear relation with the response, y. Let d be the dose level of the drug and $x = \log_e(d)$. Which of the following three models seems the most appropriate? Justify your answer.

Model 1: $y = \beta_o + \beta_1 d + \varepsilon$
Model 2: $y = \beta_o + \beta_1 d + \beta_2 d^2 + \varepsilon$
Model 3: $y = \beta_o + \beta_1 x + \varepsilon$

Med. **13.39** The following example is from the book **Residuals and Influence in Regression *(Cook and Weisberg, 1982).*** An experiment was conducted to investigate the amount of drug that is retained in the liver of a rat. In the experiment, rats were injected with a dose of a drug that was approximately proportional to the body weight of the rat. The amount of the drug injected into the rat was determined as approximately 40 mg of the drug per kilogram of body weight. After a set period of time, the rat was sacrificed, the animal's liver was weighed, and the fraction of the drug recovered in the liver was recorded. The experimenters wanted to relate the proportion of the drug in the rat's liver, y, to the explanatory variables: the body weight of the rat (gm), x_1: liver weight of the rat (gm), x_2: and relative dose level of the drug injected into the rat, x_3. The data are given here.

Case	x_1	x_2	x_3	y	Case	x_1	x_2	x_3	y
1	176	6.5	.88	.42	11	158	6.9	.80	.27
2	176	9.5	.88	.25	12	148	7.3	.74	.36
3	190	9.0	1.00	.56	13	149	5.2	.75	.21
4	176	8.9	.88	.23	14	163	8.4	.81	.28
5	200	7.2	1.00	.23	15	170	7.2	.85	.34
6	167	8.9	.83	.32	16	186	6.8	.94	.28
7	188	8.0	.94	.37	17	146	7.3	.73	.30
8	195	10	.98	.41	18	181	9.0	.90	.37
9	176	8.0	.88	.33	19	149	6.4	.75	.46
10	165	7.9	.84	.38					

a. Is there a problem with collinearity amongst the explanatory variables?
b. Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to the data. Evaluate the fit of this model.
c. Is it possible to obtain essentially the same degree of fit as in part (b) using a model without some of the explanatory variables? Which subset of the variables yields the best fit?

13.40 Refer to Exercise 13.39.

a. Are there any influential or leverage data values in the rat data?

b. Remove case 3 from the data set and repeat parts (b) and (c) from Exercise 13.39. Did removing case 3 greatly change your answers?

c. Why do you think case 3 had such a large impact on the modeling?

Engin.

13.41 The abrasive effect of a wear tester on a particular fabric was measured while the machine was run at six different speeds. Forty-eight identical 5-inch-square pieces of fabric were cut, with eight squares randomly assigned to each of the six machine speeds: 100, 120, 140, 160, 180, and 200 revolutions per minute (rev/min). The order of assignment of the squares to the machines was random, with each square tested for a 3-minute period at the appropriate machine setting. The amount of wear was measured and recorded for each square. The data appear in the accompanying table.

a. Plot the six mean wear values versus machine speed and suggest a model.

b. Fit the suggested model to the data.

c. Suggest which residual plots might be useful in checking the assumptions underlying the model.

Machine Speed (rev/min)	Wear
100	23.0, 23.5, 24.4, 25.2, 25.6, 26.1, 24.8, 25.6
120	26.7, 26.1, 25.8, 26.3, 27.2, 27.9, 28.3, 27.4
140	28.0, 28.4, 27.0, 28.8, 29.8, 29.4, 28.7, 29.3
160	32.7, 32.1, 31.9, 33.0, 33.5, 33.7, 34.0, 32.5
180	43.1, 41.7, 42.4, 42.1, 43.5, 43.8, 44.2, 43.6
200	54.2, 43.7, 53.1, 53.8, 55.6, 55.9, 54.7, 54.5

13.42 Refer to Exercise 13.41. Perform a test for lack of fit on the model you fit in Exercise 13.41.

13.43 Refer to the data of Exercise 13.41. Suppose that another variable was controlled, that the first four squares at each speed were treated with a .2 concentration of protective coating, and that the second four squares were treated with a .4 concentration of the same coating. Given that x_1 denotes the machine speed and x_2 denotes the concentration of the protective coating, fit these models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2 + \varepsilon$$

Engin.

13.44 A laundry detergent manufacturer wished to test a new product prior to market release. One area of concern was the relationship between the height of the detergent suds in a washing machine as a function of the amount of detergent added and the degree of agitation in the wash cycle. For a standard size washing machine tub filled to the full level, random assignments of different agitation levels (measured in minutes) and amounts of detergent were made and tested on the washing machine. The data are shown in the accompanying table.

a. Plot the data and suggest a model.

b. Does the assumption of normality appear to hold?

c. Fit an appropriate model.

d. Use residual plots to detect possible violations of the assumptions.

Height, y	Agitation, x_1	Amount, x_2	Height, y	Agitation, x_1	Amount, x_2
28.1	1	6	69.2	2	9
32.3	1	7	72.9	2	10
34.8	1	8	88.2	3	6
38.2	1	9	89.3	3	7
43.5	1	10	94.1	3	8
60.3	2	6	95.7	3	9
63.7	2	7	100.6	3	10
65.4	2	8			

13.45 Refer to Exercise 13.44. Would the following model be more appropriate? Why or why not?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \beta_6 x_1 x_2^2 + \beta_7 x_1^2 x_2 + \beta_8 x_1^2 x_2^2 + \varepsilon$$

13.46 Refer to the data of Exercise 13.44.

a. Can we test for lack of fit for the following model?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \beta_6 x_1 x_2^2 + \beta_7 x_1^2 x_2 + \beta_8 x_1^2 x_2^2 + \varepsilon$$

b. Write the complete model for the sample data. Note that if there was replication at one or more design points, the number of degrees of freedom for SS_{Lack} would be identical to the difference between the number of parameters in the complete model and the number of parameters in the model of part (a).

13.47 Refer to Example 13.1.

a. Use a variable selection procedure to determine a model for this study.

b. Do the model conditions appear to be valid for the model constructed in part (a)? Justify your answer.

c. Use your fitted model to predict the value of EHg for a lake having Alk = 80, pH = 6, Ca = 60, and Chlo = 40.

13.48 The solubility of a solution was examined for six different temperature settings, shown in the accompanying table.

y, Solubility by Weight	*x*, Temperature (°C)
43, 45, 42	0
32, 33, 37	25
21, 28, 29	50
15, 14, 9	75
12, 10, 8	100
7, 6, 2	125

a. Plot the data, and fit as appropriate.

b. Test for lack of fit if possible. Use $\alpha = .05$.

c. Examine the residuals and draw conclusions.

13.49 Refer to Exercise 13.48. Suppose we are missing the following observations: $y = 33, 28, 10$.

a. Fit the model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$.

b. Test for lack of fit, using $\alpha = .05$.

c. Again examine the residuals.

13.50 Refer to Exercise 13.41.

a. Test for lack of fit of the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2 + \varepsilon$$

b. Write the complete model for this experimental situation.

13.51 Refer to the data of Exercise 13.37. Test for lack of fit of a quadratic model.

Psy. **13.52** A psychologist wants to examine the effects of sleep deprivation on a person's ability to perform simple arithmetic tasks. To do this, prospective subjects are screened to obtain individuals whose daily sleep patterns were closely matched. From this group, 20 subjects are chosen. Each individual selected is randomly assigned to one of five groups, four individuals per group.

Group 1: 0 hours of sleep
Group 2: 2 hours of sleep
Group 3: 4 hours of sleep
Group 4: 6 hours of sleep
Group 5: 8 hours of sleep

All subjects are then placed on a standard routine for the next 24 hours.

The following day after breakfast, each individual is tested to determine the number of arithmetic additions done correctly in a 10-minute period. That evening the amount of sleep each person is allowed depends on the group to which he or she had been assigned. The following morning after breakfast, each person is again tested using a different but equally difficult set of additions.

Let the response of interest be the number of correct responses on the first test day minus the number correct on the second test day. The data are presented here.

Group	Response, y
1	39, 33, 41, 40
2	25, 29, 34, 26
3	10, 18, 14, 17
4	4, 6, −1, 9
5	−5, 0, −3, −8

a. Plot the sample data and use the plot to suggest a model.
b. Fit the suggested model.
c. Examine the fitted model for possible violation of assumptions.

Engin. **13.53** An experiment was conducted to determine the relationship between the amount of warping y for a particular alloy and the temperature (in °C) under which the experiment was conducted. The sample data appear in the accompanying table. Note that three observations were taken at each temperature setting.

Amount of Warping	Temperature (°C)
10, 13, 12	15
14, 12, 11	20
14, 12, 16	25
18, 19, 22	30
25, 21, 20	35
23, 25, 26	40
30, 31, 34	45
35, 33, 38	50

a. Plot the data to determine whether a linear or quadratic model appears more appropriate.
b. Fit a linear model and display the prediction equation. Superimpose the prediction equation over the scatter diagram of y versus x.
c. Fit a quadratic model and display the prediction equation. Superimpose the quadratic prediction equation on the scatter diagram. Which fit looks better, the linear or the quadratic?
d. Predict the amount of warping at a temperature of 27°C, using both the linear and the quadratic prediction equations.

Sci. **13.54** A soil scientist wants to relate the daily evaporation from the soil to air temperature, relative humidity, and wind speed. The scientist collects data at a number of locations in Texas on the variables maximum, minimum, and average soil temperature (x_1, x_2, x_3); maximum, minimum, and average air temperature (x_4, x_5, x_6); maximum, minimum, and average relative humidity (x_7, x_8, x_9); and total wind (x_{10}). The response is the daily amount of evaporation from the soil (y). The data are given below.

Obs	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1	84	65	147	85	59	151	95	40	398	273	30
2	84	65	149	86	61	159	94	28	345	140	34
3	84	66	142	83	64	152	94	41	388	318	33
4	79	67	147	83	65	158	94	50	406	282	26
5	81	68	167	88	69	180	93	46	379	311	41
6	74	66	131	77	67	147	96	73	478	446	4
7	73	66	131	78	69	159	96	72	462	294	5
8	75	67	134	84	68	159	95	70	464	313	20
9	84	68	161	89	71	195	95	63	430	455	31
10	86	72	169	91	76	206	93	56	406	604	38
11	88	73	178	91	76	208	94	55	393	610	43
12	90	74	187	94	76	211	94	51	385	520	47
13	88	72	171	94	75	211	96	54	405	663	45
14	88	72	171	92	70	201	95	51	392	467	45
15	81	69	154	87	68	167	95	61	448	184	11
16	79	68	149	83	68	162	95	59	436	177	10
17	84	69	160	87	66	173	95	42	392	173	30
18	84	70	160	87	68	177	94	44	392	76	29
19	84	70	168	88	70	169	95	48	398	72	23
20	77	67	147	83	66	170	97	60	431	183	16
21	87	67	166	92	67	196	96	44	379	76	37
22	89	69	171	92	72	199	94	48	393	230	50
23	89	72	180	94	72	204	95	48	394	193	36
24	93	72	186	92	73	201	94	47	386	400	54
25	93	74	188	93	72	206	95	47	389	339	44
26	94	75	199	94	72	208	96	45	370	172	41
27	93	74	193	95	73	214	95	50	396	238	45
28	93	74	196	95	70	210	96	45	380	118	42
29	96	75	198	95	71	207	93	40	365	93	50
30	95	76	202	95	69	202	93	39	357	269	48
31	84	73	173	96	69	173	94	58	418	128	17
32	91	71	170	91	69	168	94	44	420	423	20
33	88	72	179	89	70	189	93	50	399	415	15
34	89	72	179	95	71	210	98	46	389	300	42
35	91	72	182	96	73	208	95	43	384	193	44
36	92	74	196	97	75	215	96	46	389	195	41
37	94	75	192	96	69	198	95	36	380	215	49
38	96	75	195	95	67	196	97	24	354	185	53
39	93	76	198	94	75	211	93	43	364	466	53
40	88	74	188	92	73	198	95	52	405	399	21
41	88	74	178	90	74	197	95	61	447	232	1
42	91	72	175	94	70	205	94	42	380	275	44
43	92	72	190	95	71	209	96	44	379	166	44
44	92	73	189	96	72	208	93	42	372	189	46
45	94	75	194	95	71	208	93	43	373	164	47
46	96	76	202	96	71	208	94	40	368	139	50

a. Fit the following model to the data and display the fitted model.

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \varepsilon$$

b. Produce a 95% confidence interval on the average evaporation for the following values of the explanatory variables:

$x_1 = 90, x_2 = 70, x_3 = 150, x_4 = 85, x_5 = 65$

$x_6 = 180, x_7 = 95, x_8 = 40, x_9 = 375, x_{10} = 450$

13.55 Refer to Exercise 13.54.

a. Is there a strong correlation between any of the pairs of explanatory variables? What problems may result if several of the explanatory variables are highly correlated?

b. Evaluate whether the conditions of normality and equal variance hold for your model in Exercise 13.54.

13.56 Refer to Exercise 13.54.

a. Formulate a new model using a variable selection procedure with AIC and then BIC as the criterion to select the independent variables. Was there a large difference in the selected variables for the two methods?

b. Compare the standard errors of the estimated βs in the model selected by BIC to those in the full model fit in Exercise 13.54. Was there an increase or a decrease in the standard errors of the estimated βs?

c. Produce a 95% confidence interval on the average evaporation for the values of the explanatory variables given in Exercise 13.54. Was there a large difference in the two point estimators? Compare the widths of the two intervals.

13.57 Refer to Exercise 13.54. The agronomist is concerned that there may be a distinct difference between the models for land in West Texas and for land in East Texas. Observations 1–23 are data values from East Texas and 24-46 are from West Texas.

a. At the $\alpha = .05$ level, are there differences between the models for the two regions?

b. For each of the two regions, produce a 95% confidence interval on the average evaporation for the values of the explanatory variables given in Exercise 13.54.

c. Was the there a large difference in the point estimators for the two regions? Compare the widths of the intervals for the two regions.

Eco.

13.58 A random sample of 22 residential properties was used in a regression of price on nine different independent variables. The variables used in this study were as follows:

PRICE = selling price (dollars)
BATHS = number of baths (powder room = 1/2 bath)
BEDA = dummy variable for number of bedrooms (1 = 2 bedrooms, 0 = otherwise)
BEDB = dummy variable for number of bedrooms (1 = 3 bedrooms, 0 = otherwise)
BEDC = dummy variable for number of bedrooms (1 = 4 bedrooms, 0 = otherwise)
CARA = dummy variable for type of garage (1 = no garage, 0 = otherwise)
CARB = dummy variable for type of garage (1 = one-car garage, 0 = otherwise)
AGE = age in years
LOT = lot size in square yards
DOM = days on the market

In this study, homes had two, three, four, or five bedrooms and either no garage or one- or two-car garages. Hence, we are using two dummy variables to code for the three categories of garage.

Fit a full regression model (nine independent variables), and then estimate the average difference in selling price between

a. Properties with no garage and properties with a one-car garage.

b. Properties with a one-car garage and properties with a two-car garage.

c. Properties with no garage and properties with a two-car garage.

Property	PRICE	BATHS	BEDA	BEDB	BEDC	CARA	CARB	AGE	LOT	DOM
1	25750	1.0	1	0	0	1	0	23	9680	164
2	37950	1.0	0	1	0	0	1	7	1889	67
3	46450	2.5	0	1	0	0	0	9	1941	315
4	46550	2.5	0	0	1	1	0	18	1813	61
5	47950	1.5	1	0	0	0	1	2	1583	234
6	49950	1.5	0	1	0	0	0	10	1533	116
7	52450	2.5	0	0	1	0	0	4	1667	162
8	54050	2.0	0	1	0	0	1	5	3450	80

Property	PRICE	BATHS	BEDA	BEDB	BEDC	CARA	CARB	AGE	LOT	DOM
9	54850	2.0	0	1	0	0	0	5	1733	63
10	52050	2.5	0	1	0	0	0	5	3727	102
11	54392	2.5	0	1	0	0	0	7	1725	48
12	53450	2.5	0	1	0	0	0	3	2811	423
13	59510	2.5	0	1	0	0	1	11	5653	130
14	60102	2.5	0	1	0	0	0	7	2333	159
15	63850	2.5	0	0	1	0	0	6	2022	314
16	62050	2.5	0	0	0	0	0	5	2166	135
17	69450	2.0	0	1	0	0	0	15	1836	71
18	82304	2.5	0	0	1	0	0	8	5066	338
19	81850	2.0	0	1	0	0	0	0	2333	147
20	70050	2.0	0	1	0	0	0	4	2904	115
21	112450	2.5	0	0	1	0	0	1	2930	11
22	127050	3.0	0	0	1	0	0	9	2904	36

13.59 Refer to Exercise 13.58. Conduct a test using the full regression model to determine whether the depreciation (decrease) in house price per year of age is less than $2,500. Give the null hypothesis for your test and the p-value. Draw a conclusion. Use $\alpha = .05$.

13.60 Refer to Exercise 13.58. Suppose that we wished to modify our nine-variable model to allow for the possibility that the relationship between PRICE and AGE differs depending on the number of bedrooms.

a. Formulate such a model.

b. What combination of model parameters represents the difference between a five-bedroom, one-garage home and a two-bedroom, two-garage home?

13.61 Refer to Exercise 13.58. What is your choice of a "best" model from the original set of nine variables? Why did you choose this model?

13.62 Refer to Exercise 13.58. In another study involving the same 22 properties, PRICE was regressed on a single independent variable, LIST, which was the listing price of the property in thousands of dollars.

Property	PRICE	LIST
1	25750	29900
2	37950	39900
3	46450	44900
4	46550	47500
5	47950	49900
6	49950	49900
7	52450	53000
8	54050	54900
9	54850	54900
10	52050	55900
11	54392	55900
12	53450	56000
13	59510	62000
14	60102	62500
15	63850	63900
16	62050	66900
17	69450	72500
18	82304	82254
19	81850	82900
20	70050	99900
21	112450	117000
22	127050	139000

a. Fit a regression model and predict the selling price of a home that is listed at $70,000.

b. What is the chance that your prediction is off by more than $3,000?

13.63 Refer to Exercise 13.58, examine the relationship between the selling price (in thousands of dollars) of a home and two independent variables, the number of rooms and the number of square feet. Use the following data.

Property	Price	Rooms	Square Feet
1	25.75	5	986
2	37.95	5	998
3	46.45	7	1,690
4	46.55	8	1,829
5	47.95	6	1,186
6	49.95	6	1,734
7	52.45	7	1,684
8	54.05	7	1,846
9	54.85	7	1,690
10	52.05	7	1,910
11	54.39	7	1,784
12	53.45	6	1,690
13	59.51	7	1,590
14	60.10	8	1,855
15	63.85	8	2,212
16	62.05	10	2,784
17	69.45	7	2,190
18	82.30	8	2,259
19	81.85	7	1,919
20	70.05	7	1,685
21	112.45	10	2,654
22	127.05	10	2,756

a. Conduct a test to see whether the variables ROOMS and SQUARE FEET, taken together, contain information about PRICE. Use $\alpha = .05$.

b. Conduct a test to see whether the coefficient of ROOMS is equal to 0. Use $\alpha = .05$.

c. Conduct a test to see whether the coefficient of SQUARE FEET is equal to 0. Use $\alpha = .05$.

13.64 Refer to Exercise 13.63.

a. Explain the apparent inconsistency between the result of part (a) and the results of parts (b) and (c).

b. What do you think would happen to the t-value of SQUARE FEET if ROOMS was dropped from the model?

Med. **13.65** A study was conducted to determine whether infection surveillance and control programs have reduced the rates of hospital-acquired infection in U.S. hospitals. This data set consists of a random sample of 28 hospitals selected from 338 hospitals participating in a larger study. Each line of the data set provides information on variables for a single hospital. The variables are as follows:

RISK = output variable, average estimated probability of acquiring infection in hospital (in percent)

STAY = input variable, average length of stay of all patients in hospital (in days)

AGE = input variable, average age of patients (in years)

INS = input variable, ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection (times 100)

SCHOOL = dummy input variable for medical school affiliation, 1 = yes, 0 = no

RC1 = dummy input variable for region of country, 1 = northeast, 0 = other

RC2 = dummy input variable for region of country, 1 = north central, 0 = other

RC3 = dummy input variable for region of country, 1 = south, 0 = other

(Note that there are four geographic regions of the country—northeast, north central, south, and west. These four regions of the country require only three dummy variables to code for them.) The data were analyzed using SAS with the following results.

DATA LISTING

OBS	RISK	STAY	AGE	INS	SCHOOL	RC1	RC2	RC3
1	4.1	7.13	55.7	9.0	0	0	0	1
2	1.6	8.82	58.2	3.8	0	1	0	0
3	2.7	8.34	56.9	8.1	0	0	1	0
4	5.6	8.95	53.7	18.9	0	0	0	1
5	5.7	11.20	56.5	34.5	0	0	0	0
6	5.1	9.76	50.9	21.9	0	1	0	0
7	4.6	9.68	57.8	16.7	0	0	1	0
8	5.4	11.18	45.7	60.5	1	1	0	0
9	4.3	8.67	48.2	24.4	0	0	1	0
10	6.3	8.84	56.3	29.6	0	0	0	0
11	4.9	11.07	53.2	28.5	1	0	0	0
12	4.3	8.30	57.2	6.8	0	0	1	0
13	7.7	12.78	56.8	46.0	1	0	0	0
14	3.7	7.58	56.7	20.8	0	1	0	0
15	4.2	9.00	56.3	14.6	0	0	1	0
16	5.6	10.12	51.7	14.9	1	0	1	0
17	5.5	8.37	50.7	15.1	0	1	0	0
18	4.6	10.16	54.2	8.4	1	0	0	1
19	6.5	19.56	59.9	17.2	0	0	0	0
20	5.5	10.90	57.2	10.6	0	1	0	0
21	1.8	7.67	51.7	2.5	0	0	1	0
22	4.2	8.88	51.5	10.1	0	0	1	0
23	5.6	11.48	57.6	20.3	0	0	0	0
24	4.3	9.23	51.6	11.6	0	1	0	0
25	7.6	11.41	61.1	16.6	0	0	0	0
26	7.8	12.07	43.7	52.4	0	1	0	0
27	3.1	8.63	54.0	8.4	0	0	0	0
28	3.9	11.15	56.5	7.7	0	0	0	0

Does the set of seven input variables contain information about the output variable, RISK? Give a *p*-value for your test.

Based on the full regression model (seven input variables), can we be at least 95% certain that hospitals in the south have at least .5% higher risk of infection than hospitals in the west, all other things being equal?

13.66 Refer to Exercise 13.65.

a. Consider the following two statements:

> There is multicollinearity between region of the country and whether a hospital has a medical school.
>
> There is an interaction effect between region of the country and whether a hospital has a medical school.

What is the difference between these two statements? What evidence is needed to ascertain the truth or falsity of the statements? Is this evidence present in the accompanying output? If it is, do you think the statements are true or false?

b. Construct a model that allows for the possibility of an interaction effect between region of the country and medical school affiliation. For this model, what is the difference in intercept between a hospital in the northeast affiliated with a medical school and a hospital in the west not affiliated with one?

13.67 Refer to Exercise 13.65. Suppose that we decide to eliminate from the full model some variables that we think contribute little to explaining the output variable. What would your final choice of a model be? Why would you choose this model?

13.68 Refer to Exercise 13.65. Predict the infection risk of a patient in a medical school–affiliated hospital in the northeast, where the average stay of patients is 10 days, the average age is 64, and the routine culturing ratio is 20%. Is this prediction an interpolation or an extrapolation? How do you know?

Sci.

13.69 Thirty volunteers participated in the following experiment. The subjects took their own pulse rates (which is easiest to do by holding the thumb and forefinger of one hand on the pair of arteries on the side of the neck). They were then asked to flip a coin. If their coin came up heads, they ran in place for 1 minute. Then all subjects took their own pulse rates again. The difference in the before and after pulse rates was recorded, as were other data on subject characteristics. Fit a regression model to "explain" the pulse rate differences using the other variables as independent variables. The variables were

PULSE = difference between the before and after pulse rates
RUN = dummy variable, 1 = did not run in place, 0 = ran in place
SMOKE = dummy variable, 1 = does not smoke, 0 = smokes
HEIGHT = height in inches
WEIGHT = weight in pounds
PHYS1 = dummy variable, 1 = a lot of physical exercise, 0 = otherwise
PHYS2 = dummy variable, 1 = moderate physical exercise, 0 = otherwise

a. Perform an appropriate test to determine whether the entire set of independent variables explains a significant amount of the variability of PULSE. Draw a conclusion based on $\alpha = .01$.

b. Does multicollinearity seem to be a problem here? What is your evidence? What effect does multicollinearity have on your ability to make predictions using regression?

c. Based on the full regression model (six dependent variables), compute a point estimate of the average increase in PULSE for individuals who engaged in a lot of physical activity compared to those who engaged in little physical activity. Can we be 95% certain that the actual average increase is greater than 0?

LISTING OF DATA FOR EXERCISE 13.69

OBS	PULSE	RUN	SMOKE	HEIGHT	WEIGHT	PHYS1	PHYS2
1	−29	0	1	66	140	0	1
2	−17	0	1	72	145	0	1
3	−14	0	0	73	160	1	0
4	−22	0	0	73	190	0	0
5	−21	0	1	69	155	0	1
6	−25	0	1	73	165	0	0
7	−5	0	1	72	150	1	0
8	−9	0	1	74	190	0	1
9	−18	0	1	72	195	0	1
10	−23	0	1	71	138	0	1
11	−14	0	0	74	160	0	0
12	−21	0	1	72	155	0	1
13	8	0	0	70	153	1	0
14	−13	0	1	67	145	0	1
15	−21	0	1	71	170	1	0
16	−1	0	1	72	175	1	0
17	−16	0	0	69	175	0	1
18	−15	1	1	68	145	0	0
19	4	1	0	75	190	0	1
20	−3	1	1	72	180	1	0
21	2	1	0	67	140	0	1
22	−5	1	1	70	150	0	1
23	−1	1	1	73	155	0	1
24	−5	1	1	74	148	1	0
25	−6	1	0	68	150	0	1
26	−6	1	0	73	155	0	1
27	8	1	0	66	130	0	1
28	−1	1	1	69	160	0	1
29	−5	1	1	66	135	1	0
30	−3	1	1	75	160	1	0

13.70 Refer to Exercise 13.69.

a. Give the implied regression line of pulse-rate difference on height and weight for a smoker who did not run in place and who has engaged in little physical activity.

b. Consider the following two statements:

There is multicollinearity between the smoke variable and the physical activity dummy variables.

There is an interaction effect between the smoke variable and the physical activity dummy variables.

Is there any difference between these two statements? Explain the relationships that would exist in the data set if each of these two statements was correct.

13.71 Refer to Exercise 13.69.

a. What is your choice of a good predictive equation? Why did you choose that particular equation?

b. The model as constructed does not contain any interaction effects. Construct a model that allows for the possibility of an interaction effect between each pair of qualitative variables.

Sci. **13.72** The data for this exercise were taken from a chemical assay of calcium discussed in ***Brown, Healy, and Kearns (1981).*** A set of standard solutions is prepared, and these and the unknowns are read on a spectrophotometer in arbitrary units (y). A linear regression model is fit to the standards, and the values of the unknowns (x) are read off from this. The preparation of the standard and unknown solutions involves a fair amount of laboratory manipulation, and the actual concentrations of the standards may differ slightly from their target values, the very precise instrumentation being capable of detecting this. The target values are 2.0, 2.0, 2.5, 3.0, 3.0 mmol per liter; the "duplicates" are made up independently. The sequence of reading the standards and unknowns is repeated four times. Two specimens of each unknown are included in each assay, and the four sequences of readings are done twice, first with the flame conditions in the instrument optimized and then with a slightly weaker flame. y is the spectrophotometer reading and x is the actual mmol per liter.

The data in the following table relate to assays on the above pattern of a set of six unknowns performed by four laboratories. The standards are identified as 2.0A, 2.0B, 2.5, 3.0A, and 3.0B; the unknowns are identified as U1, U2, W1, W2, Y1, and Y2.

Laboratory/Solution	Measurements				Laboratory/Solution	Measurements			
1 W1	1,206	1,202	1,202	1,201	3 W1	1,090	1,098	1,090	1,100
1 2.0A	1,068	1,071	1,067	1,066	3 2.0A	969	975	969	972
1 W2	1,194	1,193	1,189	1,185	3 U2	1,088	1,092	1,087	1,085
1 2.0B	1,072	1,068	1,064	1,067	3 2.0B	969	960	960	966
1 U1	1,387	1,387	1,384	1,380	3 U1	1,270	1,261	1,261	1,269
1 2.5	1,333	1,321	1,326	1,317	3 2.5	1,196	1,196	1,209	1,200
1 U2	1,394	1,390	1,383	1,376	3 W2	1,261	1,268	1,270	1,273
1 3.0A	1,579	1,576	1,578	1,572	3 3.0A	1,451	1,440	1,439	1,449
1 Y1	1,478	1,480	1,473	1,466	3 Y1	1,352	1,349	1,353	1,343
1 3.0B	1,579	1,571	1,579	1,567	3 3.0B	1,439	1,433	1,433	1,445
1 Y2	1,483	1,477	1,482	1,472	3 Y2	1,349	1,353	1,349	1,355
2 W1	1,017	1,017	1,012	1,020	4 2.0A	1,122	1,117	1,119	1,120
2 2.0A	910	916	915	915	4 W2	1,256	1,254	1,256	1,263
2 W2	1,012	1,018	1,015	1,023	4 W1	1,260	1,251	1,252	1,264
2 2.0B	913	923	914	921	4 2.0B	1,122	1,110	1,111	1,116
2 U1	1,188	1,199	1,197	1,202	4 U2	1,453	1,447	1,451	1,455
2 2.5	1,129	1,148	1,136	1,147	4 2.5	1,386	1,381	1,381	1,387
2 U2	1,186	1,196	1,193	1,199	4 U1	1,450	1,446	1,448	1,457
2 3.0A	1,359	1,378	1,370	1,373	4 3.0A	1,656	1,663	1,659	1,665
2 Y1	1,263	1,280	1,280	1,279	4 Y2	1,543	1,548	1,543	1,545
2 3.0B	1,349	1,361	1,359	1,363	4 3.0B	1,658	1,658	1,661	1,660
2 Y2	1,259	1,269	1,259	1,265	4 Y1	1,545	1,546	1,548	1,544

a. Plot y versus x for the standards, one graph for each laboratory.
b. Fit the linear regression equation $y = \beta_0 + \beta_1 x + \varepsilon$ for each laboratory, and predict the value of x corresponding to the y for each of the unknowns. Compute the standard deviation of the predicted values of x based on the four predicted x-values for each of the unknowns.
c. Which laboratory appears to make better predictions of x, mmol of calcium per liter? Why?

13.73 Refer to Exercise 13.72. Suppose you average the y-values for each of the unknowns and fit the ys in the linear regression model of Exercise 13.72.

a. Do your linear regression lines change for each of the laboratories?
b. Will predictions of x change based on these new regression lines for the four laboratories? Explain.

13.74 Refer to Exercise 13.72. Using the independent variable x, suggest a single general linear model that could be used to fit the data from all four laboratories. Identify the parameters in this general linear model.

13.75 Refer to Exercise 13.74.
a. Fit the data to the model of Exercise 13.74.
b. Give separate regression models for each of the laboratories.
c. How do these regression models compare to the previous regression equations for the laboratories?
d. What advantage(s) might there be to fitting a single model rather than separate models for the laboratories?

Env.

13.76 The following data on air pollution in 41 U.S. cities are from **Biometry** ***(Sokal and Rohlf, 1981)***. The type of air pollution under study is the annual mean concentration of sulfur dioxide. The values of six explanatory variables were recorded in order to examine the variation in the sulfur dioxide concentrations. They are as follows:

y = annual mean concentration of sulfur dioxide (micrograms per cubic meter)
x_1 = average annual temperature (°F)
x_2 = number of manufacturing enterprises employing 20 or more workers
x_3 = population size (1970) census (thousands)
x_4 = average annual wind speed (mph)
x_5 = average annual precipitation (inches)
x_6 = average number of days with precipitation per year

City	y	x_1	x_2	x_3	x_4	x_5	x_6
1	10	70.3	213	582	6.0	7.05	36
2	13	61.0	91	132	8.2	48.52	100
3	12	56.7	453	716	8.7	20.66	67
4	17	51.9	454	515	9.0	12.95	86
5	56	49.1	412	158	9.0	43.37	127
6	36	54.0	80	80	9.0	40.25	114
7	29	57.3	434	757	9.3	38.89	111
8	14	68.4	136	529	8.8	54.47	116
9	10	75.5	207	335	9.0	59.80	128
10	24	61.5	368	497	9.1	48.34	115
11	110	50.6	3,344	3,369	10.4	34.44	122
12	28	52.3	361	746	9.7	38.74	121
13	17	49.0	104	201	11.2	30.85	103
14	8	56.6	125	277	12.7	30.58	82
15	30	55.6	291	593	8.3	43.11	123
16	9	68.3	204	361	8.4	56.77	113
17	47	55.0	625	905	9.6	41.31	111

(continued)

City	y	x_1	x_2	x_3	x_4	x_5	x_6
18	35	49.9	1,064	1,513	10.1	30.96	129
19	29	43.5	699	744	10.6	25.94	137
20	14	54.5	381	507	10.0	37.00	99
21	56	55.9	775	622	9.5	35.89	105
22	14	51.5	181	347	10.9	30.18	98
23	11	56.8	46	244	8.9	7.77	58
24	46	47.6	44	116	8.8	33.36	135
25	11	47.1	391	463	12.4	36.11	166
26	23	54.0	462	453	7.1	39.04	132
27	65	49.7	1,007	751	10.9	34.99	155
28	26	51.5	266	540	8.6	37.01	134
29	69	54.6	1,692	1,950	9.6	39.93	115
30	61	50.4	347	520	9.4	36.22	147
31	94	50.0	343	179	10.6	42.75	125
32	10	61.6	337	624	9.2	49.10	105
33	18	59.4	275	448	7.9	46.00	119
34	9	66.2	641	844	10.9	35.94	78
35	10	68.9	721	1,233	10.8	48.19	103
36	28	51.0	137	176	8.7	15.17	89
37	31	59.3	96	308	10.6	44.68	116
38	26	57.8	197	299	7.6	42.59	115
39	29	51.1	379	531	9.4	38.79	164
40	31	55.2	35	71	6.5	40.75	148
41	16	45.7	569	717	11.8	29.07	123

A model relating y to the six explanatory variables is of interest in order to determine which of the six explanatory variables are related to sulfur dioxide pollution and to be able to predict air pollution for given values of the explanatory variables.

- **a.** Plot y versus each of the explanatory variables. From your plots, determine if higher-order terms are needed in any of the explanatory variables.
- **b.** Is there any evidence of collinearity in the data?
- **c.** Obtain VIF for each of the explanatory variables from fitting a first-order model relating y to x_1 through x_6. Do there appear to be any collinearity problems based on the VIF values?

13.77 Refer to Exercise 13.76.

- **a.** Use a variable selection program to obtain the best four models of all possible sizes using R^2_{adj} as your criterion. Obtain values for R^2, MSE, and Cp for each of the models.
- **b.** Using the information in part (a), select the model that you think best meets the criteria of a good fit to the data and the minimum number of variables.
- **c.** Which variables were most highly related to sulfur dioxide air pollution?

13.78 Use the model you selected in Exercise 13.77 to answer the following questions.

- **a.** Do the residuals appear to have a normal distribution? Justify your answer.
- **b.** Does the condition of constant variance appear to be satisfied? Justify your answer.
- **c.** Obtain the Box–Cox transformation of this data set.

13.79 Use the model you selected in Exercise 13.77 to answer the following questions.

- **a.** Do any of the data points appear to have high influence? Leverage? Justify your answer.
- **b.** If you identified any high leverage or high influence points in part (a), compare the estimated models with and without these points.
- **c.** What is your final model describing sulfur dioxide air pollution?
- **d.** Display any other explanatory variables that may improve the fit of your model.

13.80 Use the model you selected in Exercise 13.79 to complete the following.

a. Estimate the average level of sulfur dioxide content of the air in a city having the following values for the six explanatory variables:

$$x_1 = 60 \quad x_2 = 150 \quad x_3 = 600 \quad x_4 = 10 \quad x_5 = 40 \quad x_6 = 100$$

b. Place a 95% confidence interval on your estimated sulfur dioxide level.

c. List any major limitations in your estimation of this mean.

Edu.

13.81 In Chapter 3, a data set was presented that related math and reading scores to %minority and %poverty in 22 third-, fourth-, and fifth-grade classes.

a. Fit a model that relates math scores to reading scores, %minority, and %poverty. Include two indicator variables that will allow separate slopes and intercepts for the three grade levels.

b. Test at the .05 level whether the slopes were different for the three grade levels. Interpret your results.

c. Test at the .05 level whether the intercepts were different for the three grade levels. Interpret your results.

d. Do the conditions of normality and equal variances appear to be valid for your fitted model?

e. Note that the schools had an unequal number of students in each of the three grade levels. What is the impact on your fitted regression of ignoring the size of the school?

13.82 Refer to Exercise 13.81.

a. Is reading scores, %minority, or %poverty the best predictor of math scores?

b. Estimate the average math score for a third-grade class having a reading score of 170, %minority of 40%, and %poverty of 30%. Provide both a point estimator and a 95% confidence interval.

c. Repeat the question in part (b) for both fourth- and fifth-grade classes. How different were your point estimators for the three grade levels?

13.83 Refer to Exercise 13.81.

a. Fit a second-order model relating math scores to reading scores, %minority, and %poverty with indicator variables for grade level. Does this model appear to provide a substantial improvement in fit over the first-order model?

b. Using your fitted model, estimate the average math score for a third-grade class having a reading score of 170, %minority of 40%, and %poverty of 30%. Provide both a point estimator and a 95% confidence interval.

c. Compare the estimators from the first- and second-order models.

CHAPTER 14

Analysis of Variance for Completely Randomized Designs

14.1 Introduction and Abstract of Research Study

In Section 2.5, we introduced the concepts involved in designing an experiment. It would be very beneficial to review the material in Section 2.5 prior to reading the material in Chapters 14–19. The concepts covered in Section 2.5 are fundamental to the scientific process, in which hypotheses are formulated, experiments (studies) are planned, data are collected and analyzed, and conclusions are reached, which, in turn, leads to the formulation of new hypotheses. To obtain logical conclusions from the experiments (studies), it is mandatory that the hypotheses be precisely and clearly stated and that the experiments be carefully designed, appropriately conducted, and properly analyzed. The analysis of a designed experiment requires the development of a model of the physical setting and a clear statement of the conditions under which this model is appropriate. Finally, a scientific report of the results of the experiment should contain graphical representations of the data, a verification of model conditions, a summary of the statistical analysis, and conclusions concerning the research hypotheses. In this chapter, we will discuss some standard experimental designs and their analyses.

Section 14.2 reviews the analysis of variance for a completely randomized design discussed in Chapter 8. Here the focus of interest is the comparison of treatment means. Section 14.3 introduces experiments with a factorial treatment structure

where the focus is on the evaluation of the effects of two or more independent variables (factors) on a response rather than the on comparison of treatment means, as in the designs of Section 14.2. Particular attention is given to measuring the effects of each factor alone or in combination with the other factors. Not all designs focus on either the comparison of treatment means or the examination of the effects of factors on a response. Section 14.5 deals with estimation and comparisons of the treatment means for a completely randomized design with factorial treatments. Section 14.6 describes methodology for determining the number of replications.

Abstract of Research Study: Development of a Low-Fat Processed Meat

Dietary health concerns and consumer demand for low-fat products have prompted meat companies to develop a variety of low-fat meat products. Numerous ingredients have been evaluated as fat replacements with the goal of maintaining product yields and minimizing formulation costs, while retaining acceptable palatability. The ***paper "Utilization of Soy Protein Isolate and Konjac Blends in a Low-Fat Bologna (Model System)"*** **Chin, Keeton, Longnecker, and Lamkey (1999)** describes an experiment that examines several of these issues. The researchers determined that lowering the cost of production without affecting the quality of the low-fat meat product required the substitution of nonmeat ingredients such as soy protein isolates (SPI) for a portion of the meat block. Previous experiments have demonstrated SPI's effect on the characteristics of comminuted meats, but studies evaluating SPI's effect in low-fat meat applications are limited. Konjac flour has been incorporated into processed meat products to improve gelling properties and water-holding capacity, while reducing fat content. Thus, when replacing meat with SPI, it is necessary to incorporate konjac flour into the product to maintain the high-fat characteristics of the product.

The three factors identified for study were the type of konjac blend, amount of konjac blend, and percentage of SPI substitution in the meat product. There were many other possible factors of interest, including cooking time, temperature, type of meat product, and length of curing. However, the researchers selected the commonly used levels of these factors in a commercial preparation of bologna and narrowed the study to the three most important factors. This resulted in an experiment having 12 treatments, as displayed in Table 14.1.

TABLE 14.1
Treatment design for low-fat bologna study

Treatment	Level of Blend (%)	Konjac Blend	SPI (%)
1	.5	KSS	1.1
2	.5	KSS	2.2
3	.5	KSS	4.4
4	.5	KNC	1.1
5	.5	KNC	2.2
6	.5	KNC	4.4
7	1	KSS	1.1
8	1	KSS	2.2
9	1	KSS	4.4
10	1	KNC	1.1
11	1	KNC	2.2
12	1	KNC	4.4

The objective of this study was to evaluate various types of konjac blends as a partial lean-meat replacement and to characterize their effects in a very low-fat bologna model system. Two types of konjac blends (KSS = konjac flour/starch and KNC = konjac flour/carrageenan/starch), at levels .5% and 1%, and three meat protein replacement levels with SPI (1.1, 2.2, and 4.4%) were selected for evaluation.

The experiment was conducted as a completely randomized design with a $2 \times 2 \times 3$ three-factor factorial treatment structure and three replications of the 12 treatments. There were a number of response variables measured on the 36 runs of the experiment, but we will discuss the results for the texture of the final product as measured by an Instron universal testing machine.

The researchers were interested in evaluating the relationship between the mean texture of low-fat bologna and the percentage of SPI and in comparing this relationship for the two types of konjac blends at the two set levels. We will discuss the analysis of the data in Section 14.7.

14.2 Completely Randomized Design with a Single Factor

Recall that the completely randomized design is concerned with the comparison of t population (treatment) means $\mu_1, \mu_2, \ldots, \mu_t$. We assume that there are t different populations from which we are to draw independent random samples of sizes $n_1, n_2, \ldots, n_t$, respectively. In the terminology of the design of experiments, we assume that there are $n_1 + n_2 + \cdots + n_t$ homogeneous *experimental units* (people or objects on which a measurement is made). The treatments are randomly allocated to the experimental units in such a way that n_1 units receive treatment 1, n_2 units receive treatment 2, and so on. The objective of the experiment is to make inferences about the corresponding treatment (population) means.

Consider the data for a completely randomized design as arranged in Table 14.2.

The model for a completely randomized design with t treatments and n_i observations per treatment can be written in the form

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{with} \quad \mu_i = \mu + \tau_i$$

where the terms of the model are defined as follows:

- y_{ij}: Observation on jth experimental unit receiving treatment i.
- μ_i: ith treatment mean.
- μ: Overall treatment mean, an unknown constant.
- τ_i: An effect due to treatment i, an unknown constant.
- ε_{ij}: A random error associated with the response from the jth experimental unit receiving treatment i. We require that the ε_{ij}s have a normal distribution with mean 0 and common variance σ_ε^2. In addition, the errors must be independent.

TABLE 14.2
A completely randomized design

Treatment					Mean
1	y_{11}	y_{12}	$\cdots$	y_{1n_1}	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2n_2}	$\bar{y}_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
t	y_{t1}	y_{t2}	$\cdots$	y_{tn_t}	$\bar{y}_{t\cdot}$

One problem with expressing the treatment means as $\mu_i = \mu + \tau_i$ is that we then have an overparameterized model; that is, there are only t treatment means, but we have $t + 1$ parameters: μ and $\tau_1, \tau_2, \ldots, \tau_t$. In order to obtain the least-squares estimates, it is necessary to put constraints on these sets of parameters. A widely used constraint is to set $\tau_t = 0$. Then we have exactly t parameters in our description of the t treatment means. However, this results in the following interpretation of the parameters:

$$\mu = \mu_t, \tau_1 = \mu_1 - \mu_t, \tau_2 = \mu_2 - \mu_t, \ldots, \tau_{t-1} = \mu_{t-1} - \mu_t, \tau_t = 0$$

Thus, for $i = 1, 2, \ldots, t - 1$, τ_i is comparing μ_i to μ_t. This is the parametrization used by most software programs.

The conditions given above for our model can be shown to imply that the jth recorded response from the ith treatment y_{ij} is normally distributed with mean $\mu_i = \mu + \tau_i$ and variance σ_ε^2. The ith treatment mean differs from μ_t by an amount τ_i, the treatment effect. Thus, a test of

$$H_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_t \quad \text{versus} \quad H_a\text{: Not all } \mu_i'\text{s are equal.}$$

is equivalent to testing

$$H_0\text{: } \tau_1 = \tau_2 = \cdots = \tau_t = 0 \quad \text{versus} \quad H_a\text{: Not all } \tau_i'\text{s are 0.}$$

total sum of squares

Our test statistic is developed using the idea of a partition of the **total sum of squares** (TSS) of the measurements about their mean $\bar{y}_{..} = \sum_{ij} y_{ij}$, which we defined in Chapter 8 as

$$\text{TSS} = \sum_{ij} (y_{ij} - \bar{y}_{..})^2$$

The total sum of squares is partitioned into two separate sources of variability: one due to the variability among treatments and one due to the variability among the y_{ij}s within each treatment. The second source of variability is called "error" because it accounts for the variability that is not explained by treatment differences. The **partition of TSS** can be shown to take the following form:

partition of TSS

$$\sum_{ij} (y_{ij} - \bar{y}_{..})^2 = \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{ij} (y_{ij} - \bar{y}_{i.})^2$$

When the number of replications is the same for all treatments—that is, $n_1 = n_2 = \cdots = n_t = n$—the partition becomes

$$\sum_{ij} (y_{ij} - \bar{y}_{..})^2 = n \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{ij} (y_{ij} - \bar{y}_{i.})^2$$

between-treatment sum of squares

The first term on the right side of the equal sign measures the variability of the treatment means $\bar{y}_{i.}$ about the overall mean $\bar{y}_{..}$. Thus, it is called the **between-treatment sum of squares** (SST) and is a measure of the variability in the y_{ij}s due to differences between the treatment means, μ_is. It is given by

$$\text{SST} = n \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

sum of squares for error

The second quantity is referred to as the **sum of squares for error** (SSE) and represents the variability in the y_{ij}s not explained by differences in the treatment means. This variability represents the differences in the experimental units prior to applying the treatments and the differences in the conditions that each experimental unit is exposed to during the experiment. It is given by

$$\text{SSE} = \sum_{ij} (y_{ij} - \bar{y}_{i.})^2$$

TABLE 14.3
Analysis of variance table for a completely randomized design

Source	SS	df	MS	*F*
Treatments	SST	$t-1$	$\text{MST} = \text{SST}/(t-1)$	MST/MSE
Error	SSE	$N-t$	$\text{MSE} = \text{SSE}/(N-t)$	
Total	TSS	$N-1$		

Recall from Chapter 8 that we summarized this information in an analysis of variance (AOV) table, as represented in Table 14.3, with $N = \sum_i n_i$.

unbiased estimates

When H_0: $\tau_1 = \cdots = \tau_t = 0$ is true, both MST and MSE are **unbiased estimates** of σ_ε^2, the variance of the experimental error. That is, when H_0 is true, both MST and MSE have a mean value in repeated sampling, called the **expected mean squares,** equal to σ_ε^2. We express these terms as

expected mean squares

$$E(\text{MST}) = \sigma_\varepsilon^2 \quad \text{and} \quad E(\text{MSE}) = \sigma_\varepsilon^2$$

Thus, we would expect $F = \text{MST/MSE}$ to be near 1 when H_0 is true. When H_a is true and there is a difference in the treatment means, the mean of MSE is still an unbiased estimate of σ_ε^2:

$$E(\text{MSE}) = \sigma_\varepsilon^2$$

However, MST is no longer unbiased for σ_ε^2. In fact, the expected mean square for treatments can be shown to be

$$E(\text{MST}) = \sigma_\varepsilon^2 + n\theta_T$$

where $\theta_T = \dfrac{1}{t-1}\displaystyle\sum_{i=1}^{t} n_i(\mu_i - \mu_\cdot)^2$. When H_a is true, some of the $(\mu_i - \mu_\cdot)^2$ are not zero, and θ_T is positive. Thus, MST will tend to overestimate σ_ε^2. Hence, under H_a, the ratio $F = \text{MST/MSE}$ will tend to be greater than 1, and we will reject H_0 in the upper tail of the distribution of F.

In particular, for selected values of the probability of Type I error α, we will reject H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$ if the computed value of F exceeds $F_{\alpha, t-1, N-t}$, the critical value of F found in Table 8 in the Appendix with Type I error probability α, $\text{df}_1 = t - 1$, and $\text{df}_2 = N - t$. Note that df_1 and df_2 correspond to the degrees of freedom for MST and MSE, respectively, in the AOV table.

The completely randomized design has several advantages and disadvantages when used as an experimental design for comparing t treatment means.

Advantages and Disadvantages of a Completely Randomized Design

Advantages

1. The design is extremely easy to construct.
2. The design is easy to analyze even though the sample sizes might not be the same for each treatment.
3. The design can be used for any number of treatments.

Disadvantages

1. The experimental units to which treatments are applied must be as homogeneous as possible. Any extraneous sources of variability will tend to inflate the error term, making it more difficult to detect differences among the treatment means.

As discussed in previous chapters, the statistical procedures are based on the condition that the data from an experiment constitute a random sample from a population of responses. In most cases, we have further stipulated that the population of responses have a normal distribution. When the experiment consists of randomly selected experimental units or responses from existing populations, we can in fact verify whether or not this condition is valid. However, in those experiments in which we select experimental units to meet specific criteria or the experimental units are available plots or land in an agricultural research farm, the idea that the responses from these units form a random sample from a specific population is somewhat questionable. However, in the book *The Design of Experiments,* Fisher (1966), the author demonstrated that the random assignment of treatments to experimental units provided appropriate reference populations needed for the theoretical derivation of the estimation of parameters, confidence intervals, and tests of hypotheses. That is, the random assignment of treatments to experimental units simulates the effect of independence and allows the researcher to conduct tests and estimation procedures as if the observed responses were randomly selected from an existing population.

Other justifications for randomization are based on the need to minimize biases that may arise when comparing treatments due to systematic assignments of treatments to experimental units. A researcher may subconsciously assign the "preferred" treatment to the experimental units that are more likely to produce a desired response. The technician may find it is more convenient to perform the experiments using the 10 replications of treatment T_1 in the morning, followed by the 10 replications of treatment T_2 in the afternoon. Thus, if experiments in the morning tend to provide a higher response than experiments in the afternoon, treatment T_1 would have an advantage over T_2 before the experiment was even performed.

When we are dealing with the situation in which we are randomly assigning treatments to the experimental units and then observing the responses, it is a requirement of the inference procedures discussed in this book that these observations be independent. In more advanced books, methods are available for dealing with dependent data such as time-series data or spatially correlated data. To obtain valid results, it is necessary that the observations be independently distributed. The data values are often dependent when there are physical relationships between the experimental units, such as the manner in which pots of plants are placed on a greenhouse bench, the physical proximity of test animals in a laboratory, the fact that multiple animals feed from the same container, or the location of experimental plots in a field. To minimize the possibility of experimental biases and dependency in the data and to obtain valid reference distributions, it is necessary to randomly assign the treatments to the experimental units. However, the random assignment of treatments to experimental units does not completely eliminate the problem of correlated data values. Correlation can also result from the other circumstances that may occur during the experiment. Thus, the experimenter must always be aware of any physical mechanisms that may enter the experimental setting and result in correlated responses—that is, the responses from a given experimental unit having an impact on the responses from other experimental units.

Suppose we have N homogeneous experimental units and t treatments. We want to randomly assign the ith treatment to r_i experimental units, where $r_1 + r_2 + \cdots + r_t = \text{N}$. The random assignment involves the following steps:

1. Number the experimental units from 1 to N.
2. Use a random number table or a computer program to obtain a list of numbers that is a random permutation of the numbers 1 to N.

3. Assign treatment 1 to the experimental units labeled with the first r_1 numbers in the list. Treatment 2 is assigned to the experimental units labeled with the next r_2 numbers. This process is continued until treatment t is assigned to the experimental units labeled with the last r_t numbers in the list.

We will illustrate this procedure in the Example 14.1.

EXAMPLE 14.1

An important factor in road safety on rural roads is the use of reflective paint to mark the lanes on highways. This provides lane references for drivers on roads with little or no evening lighting. A problem with the currently used paint is that it does not maintain its reflectivity over long periods of time. A researcher will be conducting a study to compare three new paints (P_2, P_3, P_4) to the currently used paint (P_1). The paints will be applied to sections of highway 6 feet in length. The response variable will be the percentage decrease in reflectivity of the markings 6 months after application. There are 16 sections of highway, and each type of paint is randomly applied to 4 sections of highway. How should the researcher assign the four paints to the 16 sections so that the assignment is completely random?

Solution Following the procedure outlined above, we number the 16 sections from 1 to 16. Next, we obtain a random permutation of the numbers 1 to 16. Using a software package, we obtain the following random permutation:

2 11 12 1 16 13 9 3 14 5 8 7 15 10 4 6

We thus obtain the assignment of paints to the highway sections as given in Table 14.4.

TABLE 14.4 Random assignments of types of paint

Section	2	11	12	1	16	13	9	3	14	5	8	7	15	10	4	6
Paint	P_1	P_1	P_1	P_1	P_2	P_2	P_2	P_2	P_3	P_3	P_3	P_3	P_4	P_4	P_4	P_4

EXAMPLE 14.2

Suppose the researcher conducts the experiment as described in Example 14.1. The reflective coating is applied to the 16 highway sections, and 6 months later the decrease in reflectivity is computed at each section. The resulting measurements are given in Table 14.5. Is there significant evidence at the $\alpha = .05$ level that the four paints have different mean reductions in reflectivity?

Solution

TABLE 14.5 Reflectivity measurements

Section	**1**	**2**	**3**	**4**	**Mean**
Paint P_1	28	35	27	21	27.75
P_2	21	36	25	18	25
P_3	26	38	27	17	27
P_4	16	25	22	18	20.25

Paint P_4 has the smallest decrease in reflectivity, so it appears to be able to maintain its reflectivity longer than the other three paints. We will now attempt to confirm this observation by testing the hypotheses

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ versus } H_a: \text{Not all } \mu_i\text{s are equal.}$$

We will construct the AOV table by computing the sum of squares using the formulas given previously:

$$\bar{y}_{..} = \frac{y_{..}}{N} = \frac{400}{16} = 25$$

$$\text{TSS} = \sum_{ij}(y_{ij} - \bar{y}_{..})^2$$

$$= (28 - 25)^2 + (35 - 25)^2 + \cdots + (22 - 25)^2 + (18 - 25)^2 = 692$$

$$\text{SST} = n\sum_{i}(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= 4[(27.75 - 25)^2 + (25 - 25)^2 + (27 - 25)^2 + (20.25 - 25)^2]$$

$$= 136.5$$

$$\text{SSE} = \text{TSS} - \text{SST} = 692 - 136.5 = 555.5$$

We can now complete the AOV table as shown in Table 14.6.

TABLE 14.6
AOV table for Example 14.2

Source	SS	df	MS	*F*	*p*-value
Treatments	136.5	3	45.5	.98	.4346
Error	555.5	12	46.292		
Total	692	15			

Because p-value $= .4346 > .05 = \alpha$, we fail to reject H_0. There is not a significant difference in the mean decreases in reflectivity for the four types of paints. ■

The researcher is somewhat concerned about the results of the study described in Example 14.2 because he was certain that at least one of the paints would show some improvement over the currently used paint. He examines the road conditions and amount of traffic flow on the 16 sections used in the study and finds that the roadways had a very low traffic volume during the study period. He decides to redesign the study to improve the generalization of the results and will include four different locations having different amounts of traffic volumes in the new study. Chapter 15 will describe how to conduct this experiment, in which we may have a second source of variability, location of the sections.

14.3 Factorial Treatment Structure

In this section, we will discuss how treatments are constructed from several factors rather than just being t levels of a single factor. These types of experiments are involved with examining the effect of two or more explanatory variables on a response variable y. For example, suppose a company has developed a new adhesive for use in the home and wants to examine the effects of temperature and humidity on the bonding strength of the adhesive. Several treatment design questions arise in any study. First, we must consider what factors (explanatory variables) are of greatest interest. Second, the number of levels and the actual settings of these levels for each of the factors must be determined. Third, having separately selected the levels for each factor, we must choose the factor–level combinations (treatments) that will be applied to the experimental units.

The ability to choose the factors and the appropriate settings for each of the factors depends on the budget, the time to complete the study, and, most important, the experimenter's knowledge of the physical situation under study. In many cases,

this will involve conducting a detailed literature review to determine the current state of knowledge in the area of interest. Then, assuming that the experimenter has chosen the levels of each independent variable, he or she must decide which factor–level combinations are of greatest interest and are viable. In some situations, certain of the factor–level combinations will not produce an experimental setting that can elicit a reasonable response from the experimental unit. Certain combinations may not be feasible due to toxicity or practicality issues.

one-at-a-time approach

As discussed in Chapter 2, one approach for examining the effects of two or more factors on a response is the **one-at-a-time approach.** To examine the effect of a single variable, an experimenter changes the levels of this variable while holding the levels of the other independent variables fixed. This process is continued for each variable while holding the other independent variables constant. Suppose that an experimenter is interested in examining the effects of two independent variables, nitrogen and phosphorus, on the yield of a crop. For simplicity, we will assume two levels of each variable have been selected for the study: 40 and 60 pounds per plot for nitrogen and 10 and 20 pounds per plot for phosphorus. For this study, the experimental units are small, relatively homogeneous plots that have been partitioned from the acreage of a farm. For our experiment, the factor–level combinations chosen might be as shown in Table 14.7. These factor–level combinations are illustrated in Figure 14.1.

From the graph in Figure 14.1, we see that there is one difference that can be used to measure the effects of nitrogen and phosphorus separately. The difference in responses for combinations 1 and 2 would estimate the effect of nitrogen; the difference in for combinations 2 and 3 would estimate the effect of phosphorus.

Hypothetical yields corresponding to the three factor–level combinations of our experiment are given in Table 14.8. Suppose the experimenter is interested in

TABLE 14.7
Factor–level combinations for a one-at-a-time approach

Combination	Nitrogen	Phosphorus
1	60	10
2	40	10
3	40	20

FIGURE 14.1
Factor–level combinations for a one-at-a-time approach

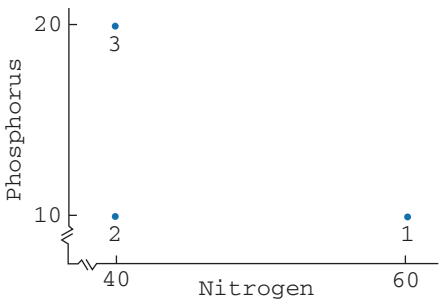


TABLE 14.8
Yields for the three factor–level combinations

Combination	Nitrogen	Phosphorus	Yield
1	60	10	145
2	40	10	125
3	40	20	160
.	60	20	?

using the sample information to determine the factor–level combination that will give the maximum yield. From the table, we see that crop yield increases when the nitrogen application is increased from 40 to 60 (holding phosphorus at 10). Yield also increases when the phosphorus setting is changed from 10 to 20 (at a fixed nitrogen setting of 40). Thus, it might seem logical to predict that increasing both the nitrogen and the phosphorus applications to the soil will result in a larger crop yield. The fallacy in this argument is that our prediction is based on the assumption that the effect of one factor is the same for both levels of the other factor.

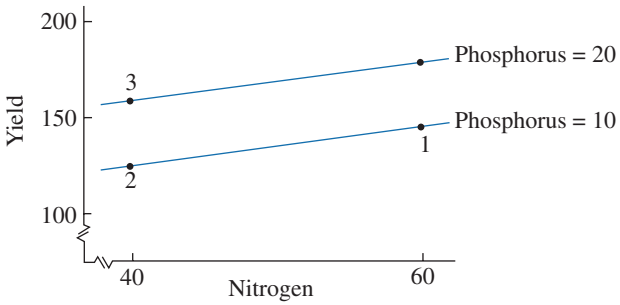
We know from our investigation what happens to yield when the nitrogen application is increased from 40 to 60 for a phosphorus setting of 10. But will the yield also increase by approximately 20 units when the nitrogen application is changed from 40 to 60 at a setting of 20 for phosphorus?

To answer this question, we could apply the factor–level combination of 60 nitrogen–20 phosphorus to another experimental plot and observe the crop yield. If the yield is 180, then the information obtained from the three factor–level combinations would be correct and would have been useful in predicting the factor–level combination that produces the greatest yield. However, suppose the yield obtained from the high settings of nitrogen and phosphorus turns out to be 110. If this happens, the two factors, nitrogen and phosphorus, are said to **interact.** That is, the effect of one factor on the response does not remain the same for different levels of the second factor, and the information obtained from the one-at-a-time approach would lead to a faulty prediction.

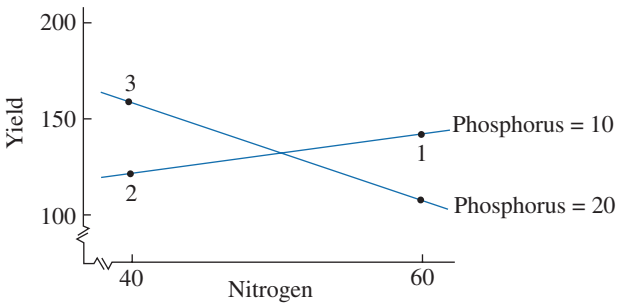
interaction

The two outcomes just discussed for the crop yield at the 60–20 setting are displayed in Figure 14.2, along with the yields at the three initial design points. Figure 14.2(a) illustrates a situation with no interaction between the two factors. The effect of nitrogen on yield is the same for both levels of phosphorus. In contrast, Figure 14.2(b) illustrates a case in which the two factors, nitrogen and phosphorus, do interact.

FIGURE 14.2
Yields of the three design points and possible yield at a fourth design point



(a) No interaction



(b) Interaction present

We have seen that the one-at-a-time approach to investigating the effect of two factors on a response is suitable only for situations in which the two factors do not interact. Although this was illustrated for the simple case in which two factors were to be investigated at each of two levels, the inadequacies of a one-at-a-time approach are even more salient when trying to investigate the effects of more than two factors on a response.

factorial treatment structures

Factorial treatment structures are useful for examining the effects of two or more factors on a response y, whether or not interaction exists. As before, the choice of the number of levels of each variable and the actual settings of these variables is important. However, assuming that we have made these selections with help from an investigator knowledgeable in the area being examined, we must decide at what factor–level combinations we will observe y.

Classically, factorial treatment structures have not been referred to as designs because they deal with the choice of levels and the selection of factor–level combinations (treatments) rather than with how the treatments are assigned to experimental units. Unless otherwise specified, we will assume that treatments are assigned to experimental units at random. The factor–level combinations will then correspond to the "treatments" of a completely randomized design.

DEFINITION 14.1

A **factorial treatment structure** is an experiment in which the response y is observed at all factor–level combinations of the independent variables.

Using our previous example, if we are interested in examining the effect of two levels of nitrogen, x_1, at 40 and 60 pounds per plot and two levels of phosphorus, x_2, at 10 and 20 pounds per plot on the yield of a crop, we could use a completely randomized design where the four factor–level combinations (treatments) of Table 14.9 are assigned at random to the experimental units.

Similarly, if we wished to examine x_1 at two levels—40 and 60—and x_2 at the three levels—10, 15, and 20—we could use the six factor–level combinations of Table 14.10 as treatments in a completely randomized design.

TABLE 14.9
2×2 factorial treatment structure for crop yield

Factor–Level Combinations		
x_1	x_2	Treatment
40	10	1
40	20	2
60	10	3
60	20	4

TABLE 14.10
2×3 factorial treatment structure for crop yield

Factor–Level Combinations		
x_1	x_2	Treatment
40	10	1
40	15	2
40	20	3
60	10	4
60	15	5
60	20	6

EXAMPLE 14.3

A horticulturist is interested in the impact of water loss due to transpiration on the yields of tomato plants. The researcher would provide covers for the tomato plants at various stages of their development. Small plots of land planted with tomatoes would be shaded to reduce the amount of sunlight to which the tomato plants were exposed. The levels of shading would be reductions of 0, 1/4, 1/2, and 3/4 in the normal sunlight that the plots naturally receive. Plant development would be divided into three stages: stage I, stage II, and stage III. Provide the factor–level combinations (treatments) to be used in a completely randomized experiment with a 3×4 factorial treatment structure.

Solution The 3×4 factor–level combinations result in 12 treatments, as displayed in Table 14.11.

TABLE 14.11
Treatments from factorial combinations

	Treatment											
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Growth stage	I	I	I	I	II	II	II	II	III	III	III	III
Shading	0	1/4	1/2	3/4	0	1/4	1/2	3/4	0	1/4	1/2	3/4

The examples of factorial treatment structures presented in this section have concerned two independent variables. However, the procedure applies to any number of factors and levels per factor. Thus, if we had four different factors—F_1, F_2, F_3, and F_4—at two, three, three, and four levels, respectively, we could formulate a $2 \times 3 \times 3 \times 4$ factorial treatment structure by considering all $2 \cdot 3 \cdot 3 \cdot 4 = 72$ factor–level combinations.

One final comparison should be made between the one-at-a-time approach and a factorial treatment structure. Not only do we get information concerning factor interactions using a factorial treatment structure, but also, when there are no interactions, we get at least the same amount of information about the effects of each individual factor using fewer observations. To illustrate this idea, let us consider the 2×2 factorial treatment structure with nitrogen and phosphorus. If there is no interaction between the two factors, the data appear as shown in Figure 14.3(a). For convenience, the data are reproduced in Table 14.12, with the four treatment combinations designated by the numbers 1 through 4. If a 2×2 factorial treatment structure is used and no interaction exists between the two factors, we can obtain two independent differences to use in examining the effects of each of the factors on the response. Thus, from Table 14.12, the differences between observations 1 and 4 and the difference between observations 2 and 3 would be used to measure the effect of phosphorus. Similarly, the difference between observations 4 and 3 and the difference between observations 1 and 2 would be used to measure the effect of the two levels of nitrogen on plot yield.

If we employed a one-at-a-time approach for the same experimental situation, it would take six observations (two observations at each of the three initial factor–level combinations shown in Table 14.12) to obtain the same number of independent differences for examining the separate effects of nitrogen and phosphorus when no interaction is present.

The model for an observation in a completely randomized design with a two-factor factorial treatment structure and $n > 1$ replications can be written in the form

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \quad \text{with} \quad \mu_{ij} = \mu + \tau_i + \beta_j + \tau\beta_{ij}$$

FIGURE 14.3
Illustrations of the absence and presence of interaction in a 2×2 factorial treatment structure: (a) Factors A and B do not interact. (b) Factors A and B interact. (c) Factors A and B interact.

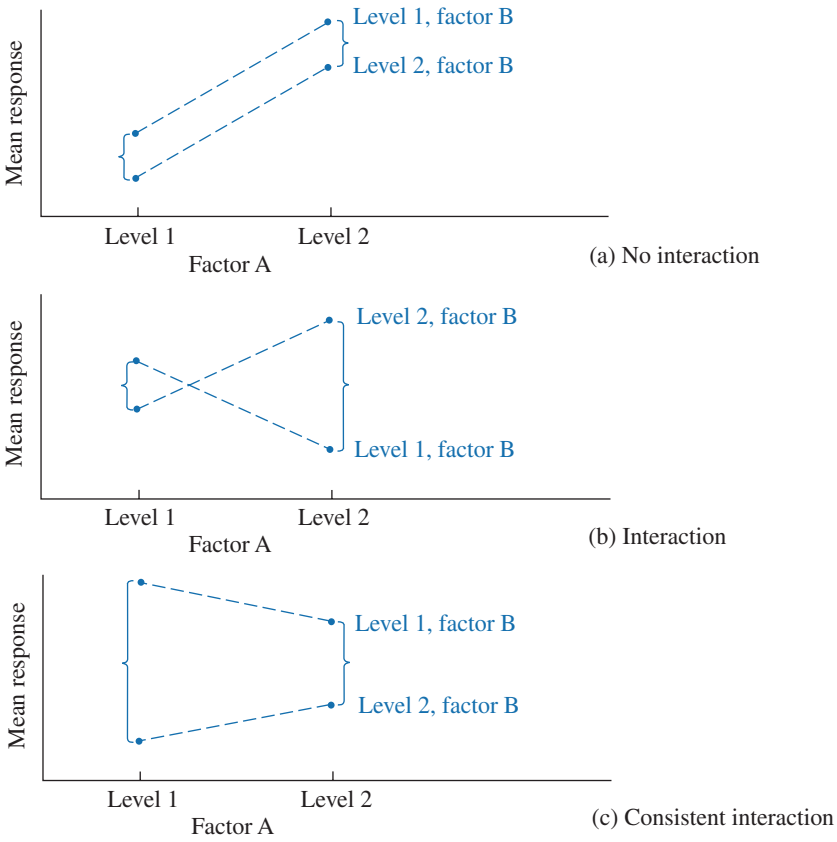


TABLE 14.12
Factor–level combinations for a 2×2 factorial treatment structure

Treatment	Nitrogen	Phosphorus	Mean Yields
1	60	10	145
2	40	10	125
3	40	20	165
4	60	20	180

where the terms of the model are defined as follows:

y_{ijk}: The response from the kth experimental unit receiving the ith level of factor A and the jth level of factor B.

μ_{ij}: (i, j) treatment mean.

μ: Overall mean, an unknown constant.

τ_i: An effect due to the ith level of factor A, an unknown constant.

β_j: An effect due to the jth level of factor B, an unknown constant.

$\tau\beta_{ij}$: An interaction effect of the ith level of factor A with the jth level of factor B, an unknown constant.

ε_{ijk}: A random error associated with the response from the kth experimental unit receiving the ith level of factor A combined with the jth level of factor B. We require that the ε_{ij}s have a normal distribution with mean 0 and common variance σ_ε^2. In addition, the errors must be independent.

TABLE 14.13
Expected values for a 2×2 factorial treatment structure without interactions

	Factor B	
Factor A	**Level 1**	**Level 2**
Level 1	$\mu + \tau_1 + \beta_1$	$\mu + \tau_1 + \beta_2$
Level 2	$\mu + \tau_2 + \beta_1$	$\mu + \tau_2 + \beta_2$

The conditions given for our model can be shown to imply that the recorded response from the kth experimental unit receiving the ith level of factor A combined with the jth level of factor B is normally distributed with mean

$$\mu_{ij} = E(y_{ijk}) = \mu + \tau_i + \beta_j + \tau\beta_{ij}$$

and variance σ_ε^2.

To illustrate this model, consider the model for a two-factor factorial treatment structure with *no interaction,* such as the 2×2 factorial experiment with nitrogen and phosphorus:

$$y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}$$

Expected values for a 2×2 factorial experiment are shown in Table 14.13.

This model assumes that the difference in population means (expected values) for any two levels of factor A is the same no matter what level of B we are considering. The same property holds when comparing two levels of factor B. For example, the difference in mean response for levels 1 and 2 of factor A is the same value, $\tau_1 - \tau_2$, no matter what level of factor B we are considering. Thus, a test for no differences among the two levels of factor A would be of the form H_0: $\tau_1 - \tau_2 = 0$. Similarly, the difference between levels of factor B is $\beta_1 - \beta_2$ for either level of factor A, and a test of no difference between the factor B means is H_0: $\beta_1 - \beta_2 = 0$. This phenomenon was also noted for the randomized block design.

interaction

If the assumption of additivity of terms in the model does not hold, then we need a model that employs terms to account for **interaction.**

The expected values for a 2×2 factorial experiment with n observations per cell are presented in Table 14.14.

As can be seen from Table 14.14, the difference in mean response for levels 1 and 2 of factor A on level 1 of factor B is

$$(\tau_1 - \tau_2) + (\tau\beta_{11} - \tau\beta_{21})$$

but for level 2 of factor B, this difference is

$$(\tau_1 - \tau_2) + (\tau\beta_{12} - \tau\beta_{22})$$

Because the difference in mean response for levels 1 and 2 of factor A is *not* the same for different levels of factor B, the model is no longer additive, and we say that the two factors interact.

TABLE 14.14
Expected values for a 2×2 factorial treatment structure with interactions

	Factor B	
Factor A	**Level 1**	**Level 2**
Level 1	$\mu + \tau_1 + \beta_1 + \tau\beta_{11}$	$\mu + \tau_1 + \beta_2 + \tau\beta_{12}$
Level 2	$\mu + \tau_2 + \beta_1 + \tau\beta_{21}$	$\mu + \tau_2 + \beta_2 + \tau\beta_{22}$

Similar to the model for t treatments, this model is grossly overparametrized. There are ab treatment means, μ_{ij}, which have been modeled by $1 + a + b + ab = (a + 1)(b + 1)$ parameters: μ; a parameters $\tau_1, \ldots, \tau_a$; b parameters $\beta_1, \ldots, \beta_b$ and ab parameters $\tau\beta_{11}, \ldots, \tau\beta_{ab}$. In order to obtain the least-squares estimators, we place the following constraints on the effect parameters:

$$\tau_a = 0,\ \beta_b = 0,\ \tau\beta_{ij} = 0 \text{ whenever } i = a \text{ and/or } j = b$$

This leaves exactly ab nonzero parameters to describe the ab treatment means, μ_{ij}.

Under the above constraints, the relationship between the parameters μ, τ_i, β_i, and $\tau\beta_{ij}$ and the treatment means $\mu_{ij} = \mu + \tau_i + \beta_j + \tau\beta_{ij}$ becomes

a. Overall mean: $\mu = \mu_{ab}$.
b. Main effects of factor A: $\tau_i = \mu_{ib} - \mu_{ab}$ for $i = 1, 2, \ldots, a - 1$.
c. Main effects of factor B: $\beta_j = \mu_{aj} - \mu_{ab}$ for $j = 1, 2, \ldots, b - 1$.
d. Interaction effects of factors A and B: $\tau\beta_{ij} = (\mu_{ij} - \mu_{ib}) - (\mu_{aj} - \mu_{ab})$.

EXAMPLE 14.4

The treatments in an experiment are constructed by crossing the levels of factor A and factor B, both of which have two levels. Relate the parameters in the model $y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$ to the treatment means, μ_{ij}.

Solution The treatment means are related to the parameters by $\mu_{ij} = \mu + \tau_i + \beta_j + \tau\beta_{ij}$. The parameter constraints—$\tau_a = 0$, $\beta_b = 0$, and $\tau\beta_{ij} = 0$ whenever $i = a$ and/ or $j = b$—imply that $\tau_2 = 0$, $\beta_2 = 0$, and $\tau\beta_{12} = \tau\beta_{21} = \tau\beta_{22} = 0$. Therefore, we have

$\mu_{22} = \mu + \tau_2 + \beta_2 + \tau\beta_{22} = \mu$, which implies that $\mu = \mu_{22}$
$\mu_{12} = \mu + \tau_1 + \beta_2 + \tau\beta_{12} = \mu + \tau_1$, which implies that $\tau_1 = \mu_{12} - \mu = \mu_{12} - \mu_{22}$
$\mu_{21} = \mu + \tau_2 + \beta_1 + \tau\beta_{21} = \mu + \beta_1$, which implies that $\beta_1 = \mu_{21} - \mu = \mu_{21} - \mu_{22}$
$\mu_{11} = \mu + \tau_1 + \beta_1 + \tau\beta_{11} = \mu_{22} + (\mu_{12} - \mu_{22}) + (\mu_{21} - \mu_{22}) + \tau\beta_{11}$, which implies that $\tau\beta_{11} = \mu_{11} - \mu_{22} - (\mu_{12} - \mu_{22}) - (\mu_{21} - \mu_{22}) = (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$ ■

DEFINITION 14.2

Two factors A and B are said to **interact** if the difference in mean responses for two levels of one factor is not constant across levels of the second factor.

In measuring the octane rating of gasoline, interaction can occur when two components of the blend are combined to form a gasoline mixture. The octane properties of the blended mixture may be quite different than would be expected by examining each component of the mixture. Interaction in this situation could have a positive or negative effect on the performance of the blend, in which case the components are said to potentiate, or antagonize, one another.

Suppose factors A and B both have two levels. In terms of the treatment means, μ_{ij}, the concept of an interaction between factors A and B is equivalent to the following:

$$\mu_{11} - \mu_{12} \neq \mu_{21} - \mu_{22}$$

The equation is just a mathematical expression of Definition 14.2. That is, the difference between the mean responses of levels 1 and 2 of factor B at level 1 of factor A is not equal to the difference between the mean responses of levels 1 and 2 of factor B at level 2 of factor A. This is what is depicted in Figures 14.3(b) and (c). In Figure 14.3(a), $\mu_{11} - \mu_{12} = \mu_{21} - \mu_{22}$, and, hence, we would conclude that factors A and B do not interact.

When testing the research hypothesis of an interaction between the mean responses of factors A and B, we have the following set of hypotheses:

H_0: no interaction between A and B versus H_a: A and B have an interaction.

In terms of the treatment means, we have

$$H_0: \mu_{11} - \mu_{12} = \mu_{21} - \mu_{22} \quad \text{versus} \quad H_a: \mu_{11} - \mu_{12} \neq \mu_{21} - \mu_{22}$$

In terms of the model parameters, μ, τ_i, β_j, $\tau\beta_{ij}$, we have

$$H_0: \tau\beta_{11} = 0 \quad \text{versus} \quad H_a: \tau\beta_{11} \neq 0$$

profile plot

We can amplify the notion of an interaction with the **profile plots** shown previously in Figure 14.3. As we see from Figure 14.3(a), when no interaction is present, the difference in the mean response between levels 1 and 2 of factor B (as indicated by the braces) is the same for both levels of factor A. However, for the two illustrations in Figures 14.3(b) and (c), we see that the difference between the levels of factor B changes from level 1 to level 2 of factor A. For these cases, we have an interaction between the two factors.

EXAMPLE 14.5

Suppose we have a completely randomized experiment with r replications of the treatments constructed by crossing factor A, having three levels, and factor B, having three levels. The model $y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$ was fit to the data. Answer the following questions:

a. After imposing the necessary constraints on the parameters—μ, τ_i, β_j, and $\tau\beta_{ij}$—interpret these parameters in terms of the treatment means, μ_{ij}.
b. State the null and alternative hypotheses for testing for an interaction in terms of the parameters μ, τ_i, β_j, and $\tau\beta_{ij}$.
c. State the null and alternative hypotheses for testing for an interaction in terms of the treatment means.
d. Provide two profile plots, one in which there is an interaction between factors A and B and one in which there is not an interaction.

Solution

a. The constraints yield $\tau_3 = 0, \beta_3 = 0, \tau\beta_{13} = 0, \tau\beta_{23} = 0, \tau\beta_{31} = 0, \tau\beta_{32} = 0,$ and $\tau\beta_{33} = 0$. This then yields the following interpretation for the parameters:

$$\mu_{33} = \mu + \tau_3 + \beta_3 + \tau\beta_{33} = \mu + 0 \Rightarrow \mu = \mu_{33}$$
$$\mu_{23} = \mu + \tau_2 + \beta_3 + \tau\beta_{23} = \mu + \tau_2 + 0 \Rightarrow \tau_2 = \mu_{23} - \mu_{33}$$
$$\mu_{13} = \mu + \tau_1 + \beta_3 + \tau\beta_{13} = \mu + \tau_1 + 0 \Rightarrow \tau_1 = \mu_{13} - \mu_{33}$$
$$\mu_{32} = \mu + \tau_3 + \beta_2 + \tau\beta_{32} = \mu + \beta_2 + 0 \Rightarrow \beta_2 = \mu_{32} - \mu_{33}$$
$$\mu_{31} = \mu + \tau_3 + \beta_1 + \tau\beta_{31} = \mu + \beta_1 + 0 \Rightarrow \beta_1 = \mu_{31} - \mu_{33}$$
$$\mu_{21} = \mu + \tau_2 + \beta_1 + \tau\beta_{21} = \mu_{33} + (\mu_{23} - \mu_{33}) + (\mu_{31} - \mu_{33}) + \tau\beta_{21}$$
$$\Rightarrow \tau\beta_{21} = (\mu_{21} - \mu_{23}) - (\mu_{31} - \mu_{33})$$
$$\mu_{12} = \mu + \tau_1 + \beta_2 + \tau\beta_{12} = \mu_{33} + (\mu_{13} - \mu_{33}) + (\mu_{32} - \mu_{33}) + \tau\beta_{12}$$
$$\Rightarrow \tau\beta_{12} = (\mu_{12} - \mu_{13}) - (\mu_{32} - \mu_{33})$$
$$\mu_{22} = \mu + \tau_2 + \beta_2 + \tau\beta_{22} = \mu_{33} + (\mu_{23} - \mu_{33}) + (\mu_{32} - \mu_{33}) + \tau\beta_{22}$$
$$\Rightarrow \tau\beta_{22} = (\mu_{22} - \mu_{23}) - (\mu_{32} - \mu_{33})$$
$$\mu_{11} = \mu + \tau_1 + \beta_1 + \tau\beta_{11} = \mu_{33} + (\mu_{13} - \mu_{33}) + (\mu_{31} - \mu_{33}) + \tau\beta_{11}$$
$$\Rightarrow \tau\beta_{11} = (\mu_{11} - \mu_{13}) - (\mu_{31} - \mu_{33})$$

From the above, we can observe that the intersection terms, $\tau\beta_{ij}$, are measuring differences in the mean responses of two levels of factor B at two levels of factor A. For example, $\tau\beta_{21}$ is comparing the differences in the mean responses of levels 1 and 3 of factor B at level 2 of factor A with the differences in the mean responses at the same levels of factor B (1 and 3) at level 3 of factor A. Thus, $\tau\beta_{21} = 0$ yields $(\mu_{21} - \mu_{23}) = (\mu_{31} - \mu_{33})$.

b. H_0: $\tau\beta_{12} = \tau\beta_{21} = \tau\beta_{22} = \tau\beta_{11} = 0$ versus H_a: $\tau\beta_{ij} \neq 0$ for at least one pair (i, j)

c. H_0: $\mu_{ij} - \mu_{ik} = \mu_{hj} - \mu_{hk}$ for all choices of (i, j, h, k) versus
H_a: $\mu_{ij} - \mu_{ik} \neq \mu_{hj} - \mu_{hk}$ for at least one choice of (i, j, h, k)
The null hypothesis is stating that all the vertical distances between any pair of lines in the profile plots are equal for all levels of factor A.

d. The two profile plots are given in Figures 14.4(a) and (b),

FIGURE 14.4(a)
Profile plot without interaction

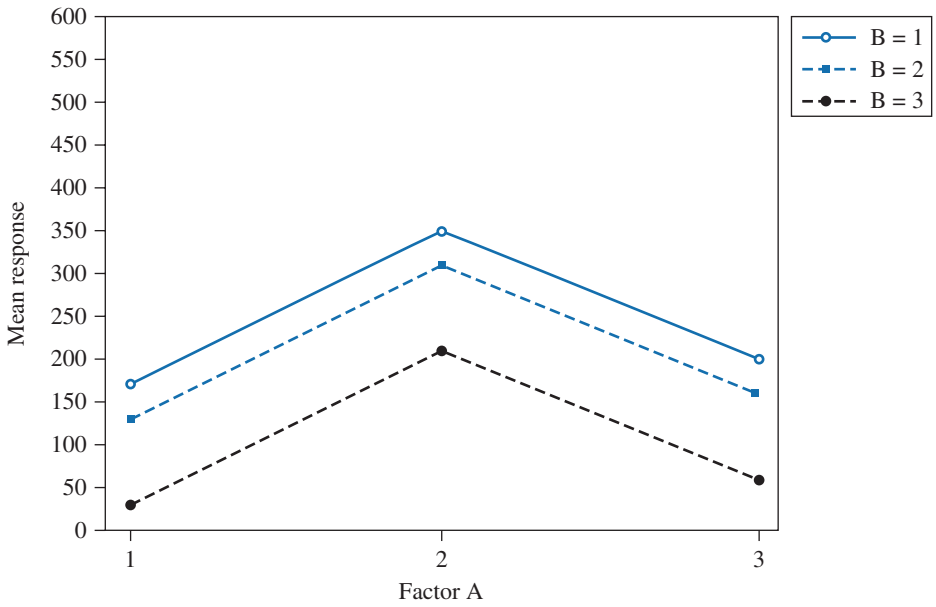
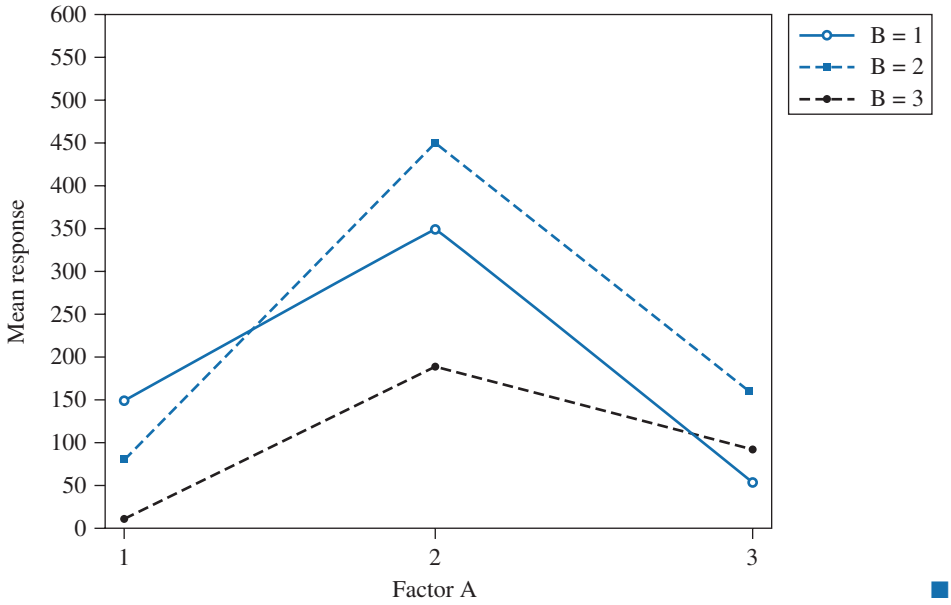


FIGURE 14.4(b)
Profile plot with interaction



Note that an interaction is not restricted to two factors. With three factors—A, B, and C—we might have an interaction between factors A and B, A and C, and B and C, and the two-factor interactions would have interpretations that follow immediately from Definition 14.2. Thus, the presence of an AC interaction indicates that the difference in mean responses for levels of factor A varies across levels of factor C. A three-way interaction among factors A, B, and C might indicate that the difference in mean responses for levels of C changes across combinations of levels for factors A and B.

The analysis of variance for a completely randomized design using a factorial treatment structure with an interaction between the factors requires that we have $n > 1$ observations on each of the treatments (factor–level combinations). We will construct the analysis of variance table for a completely randomized two-factor experiment with a levels of factor A, b levels of factor B, and n observations on each of the ab treatments. It is important to note that these results hold only when the number of replications is the *same* for all ab treatments. When the experiment has an unequal number of replications, the expressions for the sum of squares are much more complex, as will be discussed in Section 14.4. Before partitioning the total sum of squares into its components, we need the notation defined here:

y_{ijk}: Observation on the kth experimental unit receiving the ith level of factor A and jth level of factor B

$\bar{y}_{i..}$: Sample mean for observations at the ith level of factor A,

$$\bar{y}_{i..} = \frac{1}{bn}\sum_{jk} y_{ijk}$$

$\bar{y}_{.j.}$: Sample mean for observations at the jth level of factor B,

$$\bar{y}_{.j.} = \frac{1}{an}\sum_{ik} y_{ijk}$$

$\bar{y}_{ij.}$: Sample mean for observations at the ith level of factor A and the jth level of factor B, $\bar{y}_{ij.} = \frac{1}{n}\sum_{k} y_{ijk}$

$\bar{y}_{...}$: Overall sample mean, $\bar{y}_{...} = \frac{1}{abn}\sum_{ijk} y_{ijk}$

total sum of squares

The **total sum of squares** of the measurements about their mean $\bar{y}_{...}$ is defined as before:

$$\text{TSS} = \sum_{ijk}(y_{ijk} - \bar{y}_{...})^2$$

This sum of squares will be partitioned into four sources of variability: two due to the main effects of factors A and B, one due to the interaction between factors A and B, and one due to the variability from all sources not accounted for by the main effects and interaction. We call this source of variability **error.**

error

It can be shown algebraically that TSS takes the following form:

$$\sum_{ijk}(y_{ijk} - \bar{y}_{...})^2 = bn\sum_{i}(\bar{y}_{i..} - \bar{y}_{...})^2 + an\sum_{j}(\bar{y}_{.j.} - \bar{y}_{...})^2 + n\sum_{ij}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{ijk}(y_{ijk} - \bar{y}_{ij.})^2$$

We will interpret the terms in the partition using the parameter estimates. The first quantity on the right-hand side of the equal sign measures the **main effect of factor A** and can be written as

main effect of factor A

$$\text{SSA} = bn\sum_{i}(\bar{y}_{i..} - \bar{y}_{...})^2$$

TABLE 14.15
AOV table for a completely randomized two-factor factorial treatment structure

Source	SS	df	MS	*F*
Main effect				
A	SSA	$a-1$	$\text{MSA} = \text{SSA}/(a-1)$	MSA/MSE
B	SSB	$b-1$	$\text{MSB} = \text{SSB}/(b-1)$	MSB/MSE
Interaction				
AB	SSAB	$(a-1)(b-1)$	$\text{MSAB} = \text{SSAB}/(a-1)(\text{b}-1)$	MSAB/MSE
Error	SSE	$ab(n-1)$	$\text{MSE} = \text{SSE}/ab(n-1)$	
Total	TSS	$abn-1$		

main effect of factor B

SSA is a comparison of the factor A means, $\bar{y}_{i..}$, to the overall mean, $\bar{y}_{...}$. Similarly, the second quantity on the right-hand side of the equal sign measures the **main effect of factor B** and can be written as

$$\text{SSB} = an\sum_{j}(\bar{y}_{.j.} - \bar{y}_{...})^2$$

interaction effect of factors A and B

SSB is a comparison of the factor B means, $\bar{y}_{.j.}$ to the overall mean $\bar{y}_{...}$. The third quantity measures the **interaction effect of factors A and B** and can be written as

$$\text{SSAB} = n\sum_{ij}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 = n\sum_{ij}[(\bar{y}_{ij.} - \bar{y}_{...}) - (\bar{y}_{i..} - \bar{y}_{...}) - (\bar{y}_{.j.} - \bar{y}_{...})]^2$$

sum of squares for error

SSAB is a comparison of treatment means, $\bar{y}_{ij.}$, after removing main effects. The final term is the **sum of squares for error,** SSE, and represents the variability in the y_{ijk}s not accounted for by the main effects and interaction effects. There are several forms for this term. Defining the residuals from the model as before, we have $e_{ijk} = y_{ijk} - \hat{\mu}_{ij} = y_{ijk} - \bar{y}_{ij.}$. Therefore,

$$\text{SSE} = \sum_{ijk}(y_{ijk} - \bar{y}_{ij.})^2 = \sum_{ijk}(e_{ijk})^2$$

Alternatively, SSE = TSS − SSA − SSB − SSAB. We summarize the partition of the sum of squares in the AOV table as given in Table 14.15.

From the AOV table, we observe that if we have only one observation on each treatment, $n = 1$, then there are 0 degrees of freedom for error. Thus, if factors A and B interact and $n = 1$, then there are no valid tests for interactions or main effects. However, if the factors do not interact, then the interaction term can be used as the error term, and we replace SSE with SSAB. However, it would be an exceedingly rare situation to run experiments with $n = 1$, since in most cases the researcher would not know prior to running the experiment whether or not factors A and B interact. Hence, in order to have valid tests for main effects and interactions, we need $n > 1$.

EXAMPLE 14.6

An experiment was conducted to determine the effects of four different pesticides on the yield of fruit from three different varieties (B_1, B_2, and B_3) of a citrus tree. Eight trees from each variety were randomly selected from an orchard. The four pesticides were then randomly assigned to two trees of each variety, and applications were made according to recommended levels. Yields of fruit (in bushels per tree) were obtained after the test period. The data appear in Table 14.16.

TABLE 14.16
Data for the 3×4 factorial treatment structure of fruit tree yield, $n = 2$ observations per treatment

	Pesticide, A			
Variety, B	**1**	**2**	**3**	**4**
1	49	50	43	53
	39	55	38	48
2	55	67	53	85
	41	58	42	73
3	66	85	69	85
	68	92	62	99

a. Write an appropriate model for this experiment.
b. Set up an analysis of variance table, and conduct the appropriate F tests of main effects and interactions using $\alpha = .05$.
c. Construct a plot of the treatment means, called a **profile plot.**

profile plot

Solution The experiment described is a completely randomized 3×4 factorial treatment structure with factor A, pesticides, having $a = 4$ levels and factor B, variety, having $b = 3$ levels. There are $n = 2$ replications of the 12 factor–level combinations of the two factors.

a. The model for a 4×3 factorial treatment structure with interaction between the two factors is

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}, \quad \text{for } i = 1, 2, 3, 4; j = 1, 2, 3; k = 1, 2$$

where μ is the overall mean yield per tree, τ_is and β_js are main effects, and $\tau\beta_{ij}$s are interaction effects.

b. In most experiments, we would strongly recommend using a computer software program to obtain the AOV table, but to illustrate the calculations, we will construct the AOV for this example using the definitions of the individual sums of squares. To accomplish this, we use the treatment means given in Table 14.17.

TABLE 14.17
Sample means for factor–level combinations (treatments) of A and B

	Pesticide, A				
Variety, B	**1**	**2**	**3**	**4**	**Variety Means**
1	44	52.5	40.5	50.5	46.875
2	48	62.5	47.5	79	59.25
3	67	88.5	65.5	92	78.25
Pesticide means	53	67.83	51.17	73.83	61.46

We next calculate the total sum of squares. Because of rounding errors, the values for TSS, SSA, SSB, SSAB, and SSE are somewhat different from the values obtained from a computer program.

$$\text{TSS} = \sum_{ijk} (y_{ijk} - \bar{y}_{...})^2 = (49 - 61.46)^2 + (50 - 61.46)^2 + \ldots + (99 - 61.46)^2 = 7{,}187.96$$

The main effect sums of squares are

$$\text{SSA} = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2$$
$$= (3)(2)[(53 - 61.46)^2 + (67.83 - 61.46)^2 + (51.17 - 61.46)^2 + (73.83 - 61.46)^2] = 2{,}226.29$$

$$\text{SSB} = an\sum_{j=1}^{b}(\bar{y}_{j.} - \bar{y}_{...})^2$$
$$= (4)(2)[(46.875 - 61.46)^2 + (59.25 - 61.46)^2 + (78.25 - 61.46)^2] = 3{,}996.08$$

The interaction sum of squares is

$$\text{SSAB} = n\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij.} - \bar{y}_{i..} - \bar{y}_{j.} + \bar{y}_{...})^2$$
$$= (2)[(44 - 53 - 46.875 + 61.46)^2 + (48 - 53 - 59.25 + 61.46)^2 + (67 - 53 - 78.25 + 61.46)^2 + (52.5 - 67.83 - 46.875 + 61.46)^2 + (62.5 - 67.83 - 59.25 + 61.46)^2 + (88.5 - 67.83 - 78.25 + 61.46)^2 + (40.5 - 51.17 - 46.875 + 61.46)^2 + (47.5 - 51.17 - 59.25 + 61.46)^2 + (65.5 - 51.17 - 78.25 + 61.46)^2 + (50.5 - 73.83 - 46.875 + 61.46)^2 + (79 - 73.83 - 59.25 + 61.46)^2 + (92 - 73.83 - 78.25 + 61.46)^2]$$
$$= 456.92$$

The sum of squares error is obtained as

$$\text{SSE} = \text{TSS} - \text{SSA} - \text{SSB} - \text{SSAB} = 7{,}187.96 - 2{,}226.29 - 3{,}996.08 - 456.92 = 508.67$$

The analysis of variance table for this completely randomized 4×3 factorial treatment structure with $n = 2$ replications per treatment is given in Table 14.18.

TABLE 14.18 AOV table for fruit yield experiment of Example 14.6

Source	SS	df	MS	*F*
Pesticide, A	2,226.29	3	742.10	17.51
Variety, B	3,996.08	2	1,998.04	47.13
Interaction, AB	456.92	6	76.15	1.80
Error	508.67	12	42.39	
Total	7,187.96	23		

The first test of significance *must* be to test for an interaction between factors A and B because if the interaction is significant, then the main effects *may have no interpretation.* The *F* statistic is

$$F = \frac{\text{MSAB}}{\text{MSE}} = \frac{76.15}{42.39} = 1.80$$

The computed value of *F* does not exceed the tabulated value of 3.00 for $\alpha = .05$, $df_1 = 6$, and $df_2 = 12$ in the *F* tables. Hence, we have insufficient evidence to indicate an interaction between pesticide levels and variety of trees levels.

c. We can observe this lack of interaction by constructing a profile plot. Figure 14.5 contains a plot of the sample treatment means for this experiment.

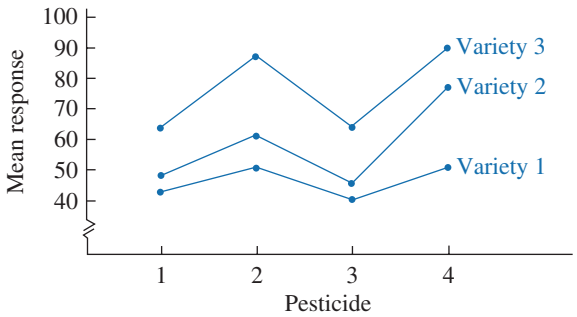


FIGURE 14.5
Profile plot for fruit yield experiment of Example 14.6

From the profile plot we can observe that the differences in mean yields among the three varieties of citrus trees remain nearly constant across the four pesticide levels. That is, the three lines for the three varieties are nearly parallel lines, and, hence, the interaction between the levels of variety and pesticide is not significant. Because the interaction is not significant, we can next test the main effects of the two factors. These tests separately examine the differences among the levels of variety and the levels of pecticides. For pesticides, the F statistic is

$$F = \frac{\text{MSA}}{\text{MSE}} = \frac{742.10}{42.39} = 17.51$$

The computed value of F does exceed the tabulated value of 3.49 for $\alpha = .05, \text{df}_1 = 3$, and $\text{df}_2 = 12$ in the F tables. Hence, we have sufficient evidence to indicate a difference in the mean yields among the four pesticide levels. For varieties, the F statistic is

$$F = \frac{\text{MSB}}{\text{MSE}} = \frac{1{,}998.04}{42.39} = 47.13$$

The computed value of F does exceed the tabulated value of 3.89 for $\alpha = .05, \text{df}_1 = 2$, and $\text{df}_2 = 12$ in the F tables. Hence, we have sufficient evidence to indicate a difference in the mean yields among the three varieties of citrus trees. ■

In Section 14.5, we will discuss how to explore which pairs of levels differ for both factors A and B.

The results of an F test for main effects for factors A or B must be interpreted very carefully in the presence of a **significant interaction.** The first thing we would do is to construct a profile plot using the sample treatment means, $\bar{y}_{ij.}$. Consider the profile plot shown in Figure 14.6. There would have been an indication of an interaction between factors A and B. Provided that the MSE was not too large relative to MSAB, the F test for interaction would undoubtedly have been significant.

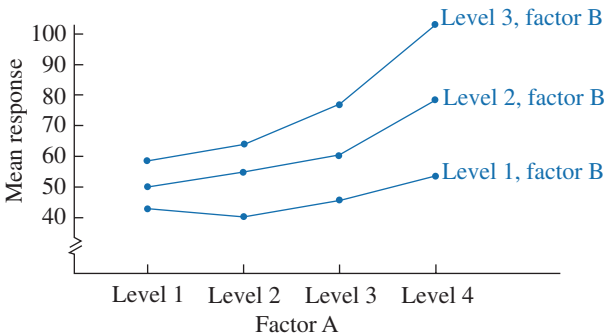
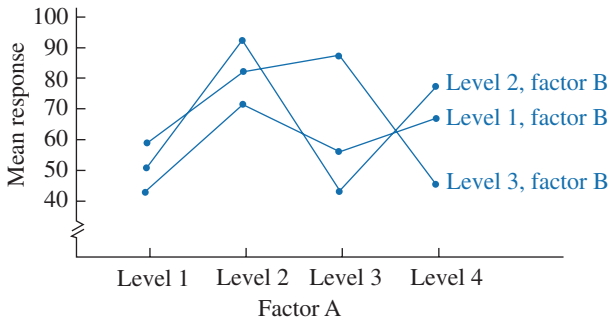


FIGURE 14.6
Profile plot in which significant interactions are present but interactions are orderly

FIGURE 14.7
Profile plot in which significant interactions are present and interactions are disorderly



Would *F* tests for main effects have been appropriate for the profile plot of Figure 14.6? The answer is no. Clearly, the profile plot in Figure 14.6 shows that the level 3 mean of factor B is always larger than the means for levels 1 and 2. Similarly, the level 2 mean for factor B is always larger than the mean for level 1 for factor B, no matter which level of factor A we examine. A significant main effect for factor B may be misleading. If we find a significant difference in the levels of factor B, with the mean response at level 3 larger than at levels 1 and 2 of factor B across all levels of factor A, we may be led to conclude that level 3 of factor B produces significantly larger mean values than the other two levels of factor B. However, note that at level 1 of factor A, there is very little difference in the mean responses for the three levels of factor B. Thus, if we were to use level 1 of factor A, the three levels of factor B would produce equivalent mean responses. As a result, our conclusions about the differences in the mean responses among the levels of factor B are not consistent across the levels of factor A and may contradict the test for main effects of factor B at certain levels of factor A.

The profile plot in Figure 14.7 shows a situation in which a test of main effects in the presence of a significant interaction might be misleading. A *disorderly* interaction, such as in Figure 14.7, can obscure the main effects. It is not that the tests are statistically incorrect; it is that they may lead to a misinterpretation of the results of the experiment. At level 1 of factor A, there is very little difference in the mean responses of the three levels of factor B. At level 3 of factor A, level 3 of factor B produces a much larger response than does level 2 of factor B. In contradiction to this result, at level 4 of factor A, level 2 of factor B produces a much larger mean response than does level 3 of factor B. Thus, when the two factors have significant interactions, conclusions about the differences in the mean responses among the levels of factor B must be made separately at *each level* of factor A. That is, a single conclusion about the levels of factor B does not hold for all levels of factor A.

When our experiment involves three factors, the calculations become considerably more complex. However, interpretations about main effects and interactions are similar to the interpretations when we have only two factors. With three factors—A, B, and C—we might have an interaction between factors A and B, A and C, and B and C. The interpretations for these two-way interactions would follow immediately from Definition 14.2. Thus, the presence of an AC interaction indicates that the differences in mean responses among the levels of factor A vary across the levels of factor C. The same care must be taken in making interpretations among main effects, as we discussed previously. A three-way interaction among factors A, B, and C might indicate that the differences in mean responses for levels of factor C change across combinations of levels for factors A and B. A

second interpretation of a three-way interaction is that the pattern in the interactions between factors A and B changes across the levels of factor C. Thus, if a three-way interaction was present and we plotted a separate profile plot for the two-way interaction between factors A and B at each level of factor C, we would see decidedly different patterns in several of the profile plots.

The model for an observation in a completely randomized design with a three-factor factorial treatment structure and $n > 1$ replications can be written in the form

$$y_{ijkm} = \mu_{ijk} + \varepsilon_{ijkm} = \mu + \tau_i + \beta_j + \gamma_k + \tau\beta_{ij} + \tau\gamma_{ik} + \beta\gamma_{jk} + \tau\beta\gamma_{ijk} + \varepsilon_{ijkm}$$

where the terms of the model are defined as follows:

y_{ijkm}: The response from the mth experimental unit receiving the ith level of factor A, the jth level of factor B, and the kth level of factor C.

μ: Overall mean, an unknown constant.

τ_i: An effect due to the ith level of factor A, an unknown constant.

β_j: An effect due to the jth level of factor B, an unknown constant.

γ_k: An effect due to the kth level of factor C, an unknown constant.

$\tau\beta_{ij}$: A two-way interaction effect of the ith level of factor A with the jth level of factor B, an unknown constant.

$\tau\gamma_{ik}$: A two-way interaction effect of the ith level of factor A with the kth level of factor C, an unknown constant.

$\beta\gamma_{jk}$: A two-way interaction effect of the jth level of factor B with the kth level of factor C, an unknown constant.

$\tau\beta\gamma_{ijk}$: A three-way interaction effect of the ith level of factor A, the jth level of factor B, and the kth level of factor C, an unknown constant.

ε_{ijkm}: A random error associated with the response from the mth experimental unit receiving the ith level of factor A combined with the jth level of factor B and the kth level of factor C. We require that the εs have a normal distribution with mean 0 and common variance σ_ε^2. In addition, the errors must be independent.

Similarly to the model with the two factors, this model is grossly overparametrized. There are abc treatment means, μ_{ijk}, which have been modeled by $1 + a + b + c + ab + ac + bc + abc = (a + 1)(b + 1)(c + 1)$ parameters: μ; a parameters $\tau_1, \ldots, \tau_a$; b parameters $\beta_1, \ldots, \beta_b$; c parameters $\gamma_1, \ldots, \gamma_c$; ab parameters $\tau\beta_{11}, \ldots, \tau\beta_{ab}$; ac parameters $\tau\gamma_{11}, \ldots, \tau\gamma_{ac}$; bc parameters $\beta\gamma_{11}, \ldots, \beta\gamma_{bc}$; and abc parameters $\tau\beta\gamma_{111}, \ldots, \tau\beta\gamma_{abc}$. In order to obtain the least-squares estimators, we need to place constraints on the effect parameters:

$$\tau_a = 0,\ \beta_b = 0,\ \gamma_c = 0$$
$$\tau\beta_{ij} = 0 \text{ whenever } i = a \text{ and/or } j = b$$
$$\tau\gamma_{ik} = 0 \text{ whenever } i = a \text{ and/or } k = c$$
$$\beta\gamma_{jk} = 0 \text{ whenever } j = b \text{ and/or } k = c$$
$$\tau\beta\gamma_{ijk} = 0 \text{ whenever } i = a \text{ and/or } j = b \text{ and/or } k = c$$

After imposing these constraints, there will be exactly abc nonzero parameters to describe the abc treatment means, μ_{ijk}.

The conditions given for our model can be shown to imply that the recorded response from the mth experimental unit receiving the ith level of factor A combined with the jth level of factor B and the kth level of factor C is normally distributed with mean

$$\mu_{ijk} = E(y_{ijkm}) = \mu + \tau_i + \beta_j + \gamma_k + \tau\beta_{ij} + \tau\gamma_{ik} + \beta\gamma_{jk} + \tau\beta\gamma_{ijk}$$

and variance σ_ε^2.

The following notation will be helpful in partitioning the total sum of squares into its components for main effects, interactions, and error.

y_{ijkm}: Observation on the mth experimental unit receiving the ith level of factor A, jth level of factor B, and kth level of factor C

$\bar{y}_{i...}$: Sample mean for observations at the ith level of factor A,

$$\bar{y}_{i...} = \frac{1}{bcn}\sum_{jkm} y_{ijkm}$$

$\bar{y}_{.j..}$: Sample mean for observations at the jth level of factor B,

$$\bar{y}_{.j..} = \frac{1}{acn}\sum_{ikm} y_{ijkm}$$

$\bar{y}_{..k.}$: Sample mean for observations at the kth level of factor C,

$$\bar{y}_{..k.} = \frac{1}{abn}\sum_{ijm} y_{ijkm}$$

$\bar{y}_{ij..}$: Sample mean for observations at the ith level of factor A and jth level of factor B,

$$\bar{y}_{ij..} = \frac{1}{cn}\sum_{km} y_{ijkm}$$

$\bar{y}_{i.k.}$: Sample mean for observations at the ith level of factor A and kth level of factor C,

$$\bar{y}_{i.k.} = \frac{1}{bn}\sum_{jm} y_{ijkm}$$

$\bar{y}_{.jk.}$: Sample mean for observations at the jth level of factor B and kth level of factor C,

$$\bar{y}_{.jk.} = \frac{1}{an}\sum_{im} y_{ijkm}$$

$\bar{y}_{ijk.}$: Sample mean for observations at the ith level of factor A, jth level of factor B, and kth level of factor C,

$$\bar{y}_{ijk.} = \frac{1}{n}\sum_{m} y_{ijkm}$$

$\bar{y}_{....}$: Overall sample mean,

$$\bar{y}_{....} = \frac{1}{abcn}\sum_{ijkm} y_{ijkm}$$

The residuals from the fitted model then become

$$e_{ijkm} = y_{ijkm} - \hat{\mu}_{ijk} = y_{ijkm} - \bar{y}_{ijk.}$$

Using the above expressions, we can partition the total sum of squares for a three-factor factorial experiment with a levels of factor A, b levels of factor B, c levels of factor C, and n observations per factor–level combination (treatments) into the sums of squares for main effects (variability between levels of a single factor), two-way interactions, a three-way interaction, and error.

The sums of squares for **main effects** are

$$\text{SSA} = bcn \sum_i (\bar{y}_{i\ldots} - \bar{y}_{\ldots})^2$$

$$\text{SSB} = acn \sum_j (\bar{y}_{.j.} - \bar{y}_{\ldots})^2$$

$$\text{SSC} = abn \sum_k (\bar{y}_{..k.} - \bar{y}_{\ldots})^2$$

The sums of squares for **two-way interactions** are

$$\text{SSAB} = cn \sum_{ij} (\bar{y}_{ij.} - \bar{y}_{\ldots})^2 - \text{SSA} - \text{SSB}$$

$$\text{SSAC} = bn \sum_{ik} (\bar{y}_{i.k.} - \bar{y}_{\ldots})^2 - \text{SSA} - \text{SSC}$$

$$\text{SSBC} = an \sum_{jk} (\bar{y}_{.jk.} - \bar{y}_{\ldots})^2 - \text{SSB} - \text{SSC}$$

The sum of squares for the **three-way interaction** is

$$\text{SSABC} = n \sum_{ijk} (\bar{y}_{ijk.} - \bar{y}_{\ldots})^2 - \text{SSAB} - \text{SSAC} - \text{SSBC} - \text{SSA} - \text{SSB} - \text{SSC}$$

The sum of squares for **error** is given by

$$\begin{aligned} \text{SSE} &= \sum_{ijkm} (e_{ijkm})^2 \\ &= \sum_{ijkm} (y_{ijkm} - \bar{y}_{ijk.})^2 \\ &= \text{TSS} - \text{SSA} - \text{SSB} - \text{SSC} - \text{SSAB} - \text{SSAC} - \text{SSBC} - \text{SSABC} \end{aligned}$$

where $\text{TSS} = \sum_{ijkm} (y_{ijkm} - \bar{y}_{\ldots})^2$.

The AOV table for a completely randomized design using a factorial treatment structure with a levels of factor A, b levels of factor B, c levels of factor C, and n observations per each of the abc treatments (factor–level combinations) is given in Table 14.19.

From the AOV table, we observe that if we have only one observation on each treatment, $n = 1$, then there are 0 degrees of freedom for error. Thus, if the interaction terms are in the model and $n = 1$, then there are no valid tests for

TABLE 14.19

AOV table for a completely randomized design with an $a \times b \times c$ factorial treatment structure

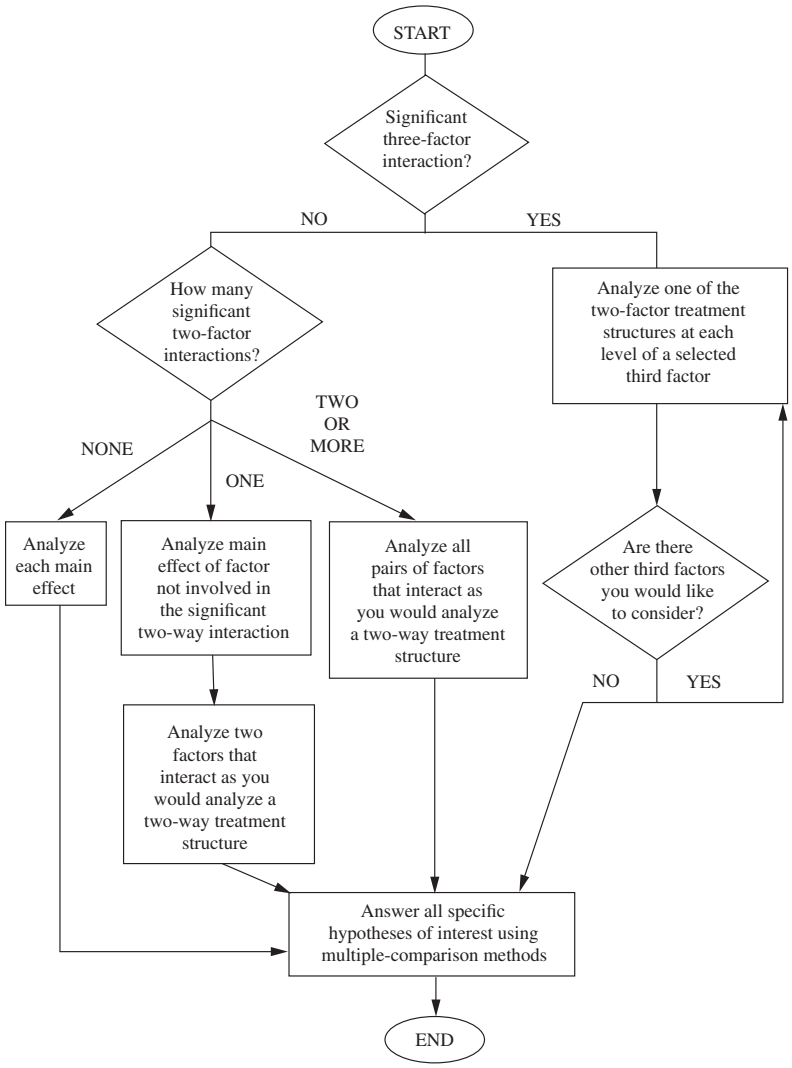
Source	SS	df	MS	*F*
Main effects				
A	SSA	$a - 1$	$\text{MSA} = \text{SSA}/(a-1)$	MSA/MSE
B	SSB	$b - 1$	$\text{MSB} = \text{SSB}/(b-1)$	MSB/MSE
C	SSC	$c - 1$	$\text{MSC} = \text{SSC}/(c-1)$	MSC/MSE
Interactions				
AB	SSAB	$(a-1)(b-1)$	$\text{MSAB} = \text{SSAB}/(a-1)(b-1)$	MSAB/MSE
AC	SSAC	$(a-1)(c-1)$	$\text{MSAC} = \text{SSAC}/(a-1)(c-1)$	MSAC/MSE
BC	SSBC	$(b-1)(c-1)$	$\text{MSBC} = \text{SSBC}/(b-1)(c-1)$	MSBC/MSE
ABC	SSABC	$(a-1)(b-1)(c-1)$	$\text{MSABC} = \text{SSABC}/(a-1)(b-1)(c-1)$	MSABC/MSE
Error	SSE	$abc(n-1)$	$\text{MSE} = \text{SSE}/abc(n-1)$	
Total	TSS	$abcn - 1$		

interactions or main effects. However, some of the interactions are known to be 0; then these interaction terms can be combined to serve as the error term in order to test the remaining terms in the model. However, it would be a rare situation to run experiments with $n = 1$ because in most cases the researcher would not know prior to running the experiment which of the interactions would be 0. Hence, in order to have valid tests for main effects and interactions, we need $n > 1$.

The analysis of a three-factor experiment is somewhat complicated by the fact that if the three-way interaction is significant, then we must handle the two-way interactions and main effects differently than when the three-way is not significant. Figure 14.8, from *Analysis of Messy Data* Vol. 1 (Milliken and Johnson, 2009), provides a general method for analyzing three-factor experiments.

We will illustrate the analysis of a three-factor experiment using Example 14.7.

FIGURE 14.8
Method for analyzing three-factor treatment structure



EXAMPLE 14.7

An industrial psychologist was studying work performance in a very noisy environment. Three factors were selected as possibly being important in explaining the variation in worker performance on an assembly line. They were noise level,

with three levels: high (HI), medium (MED), and low (LOW); gender: female (F) and male (M); and amount of experience on the assembly line: less than 5 years (E1), 5–10 years (E2), and more than 10 years (E3). Three workers were randomly selected in each of the $3 \times 2 \times 3$ factor–level combinations. We thus have a completely randomized design with a $3 \times 2 \times 3$ factorial treatment structure and three replications on each of the $t = 18$ treatments. The psychologist, process engineer, and assembly line supervisor developed a work performance index that was recorded for each of the 54 workers. The data are given in Table 14.20.

TABLE 14.20
Noise level data

Noise Level	Gender	Years of Experience	Performance Index Replication		
			y1	y2	y3
HI	F	E3	629	495	767
HI	F	E2	263	141	392
HI	F	E1	161	55	271
HI	M	E3	591	492	693
HI	M	E2	321	212	438
HI	M	E1	147	79	273
MED	F	E3	324	213	478
MED	F	E2	213	106	362
MED	F	E1	158	36	293
MED	M	E3	1,098	1,002	1,156
MED	M	E2	708	580	843
MED	M	E1	495	376	612
LOW	F	E3	1,037	902	1,183
LOW	F	E2	779	625	921
LOW	F	E1	596	458	732
LOW	M	E3	1,667	1,527	1,793
LOW	M	E2	1,192	1,005	1,306
LOW	M	E1	914	783	1,051

Use the data to determine the effect of the three factors on the mean work performance index. Use $\alpha = .05$ in all tests of hypotheses.

Solution The first step in the analysis is to examine the AOV table from the following SAS output and produce profile plots.

```
The GLM Procedure

Dependent Variable: C PERFORMANCE

                              Sum of          Mean
Source            DF         Squares        Square   F Value   Pr > F
Model             17    8963323.704     527254.336     33.22   <.0001
Error             36     571427.333      15872.981
Corrected Total   53    9534751.037

                                                     Mean
Source            DF    Type III SS        Square   F Value   Pr > F
N                  2    4460333.593    2230166.796    140.50   <.0001
G                  1    1422364.741    1422364.741     89.61   <.0001
G*N                2     689478.481     344739.241     21.72   <.0001
E                  2    2102606.259    1051303.130     66.23   <.0001
N*E                4      75059.852      18764.963      1.18   0.3351
G*E                2     114623.593      57311.796      3.61   0.0372
N*G*E              4      98857.185      24714.296      1.56   0.2068
```

From the AOV table, the p-value for the three-way interaction is .2068, which is considerably larger than $\alpha = .05$; therefore, we fail to reject the null hypothesis of no three-way interaction. Because the three-way interaction was not significant, we now consider the three two-way interactions. The interaction of noise with gender has p-value $< .0001 < .05$, which implies very significant evidence of an interaction. The interaction of noise with experience has p-value $= .3351 > .05$, which implies no significant evidence of an interaction. The interaction of gender with experience has p-value $= .0372 < .05$, which implies significant evidence of an interaction. In order to investigate the relationship among the three factors, the tables of mean responses will be presented here. Because the three-way interaction was not significant, only the two-way means will be reported in Table 14.21.

TABLE 14.21
Two-way treatment means for noise data

	Noise Level			Experience		
Gender	**Low**	**Medium**	**High**	**E1**	**E2**	**E3**
Female	803.7	242.6	352.7	306.7	422.4	669.8
Male	1,248.7	763.3	360.7	525.6	733.9	1,113.2

	Noise Level		
Experience	**Low**	**Medium**	**High**
E1	755.7	328.3	164.3
E2	971.3	468.7	294.5
E3	1,351.5	711.8	611.2

In order to confirm the lack of a three-way interaction in the three factors, the profile plots of experience by noise level, first for males and then for females, are given in Figure 14.9. The two plots are remarkably similar, except that the E3 line for females has an increase in its mean index when the noise level goes from

FIGURE 14.9(a)
Profile plot of experience by noise level for males

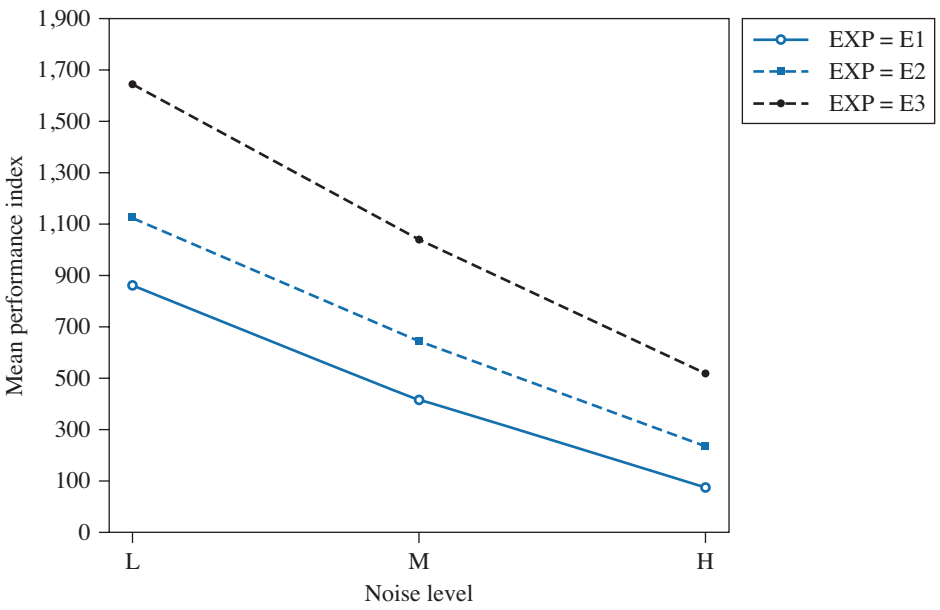
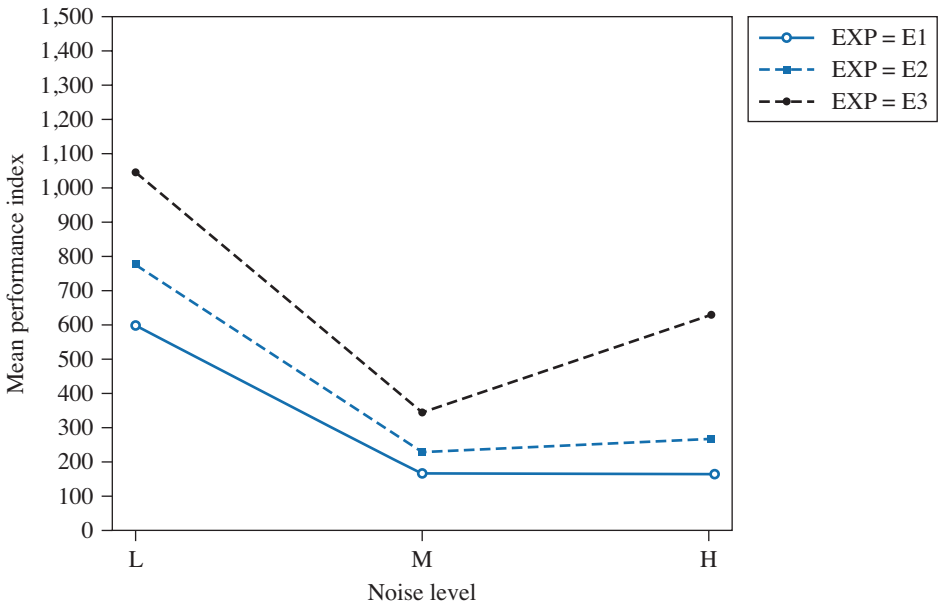


FIGURE 14.9(b)
Profile plot of experience by noise level for females



medium to high, whereas the E3 line for males, displays a decrease in its mean index. However, after taking into account the standard errors in the estimation of the treatment means, $SE(\hat{\mu}_{ij}) = 72.7$, the graphs tend to confirm the conclusion of no significant interaction that was obtained from the AOV F test. If there would have been a three-way interaction, then the relationships between the three lines in the plot for males would have been different than the plot for females. The three two-way profile plots are given in Figure 14.10. From the profile plot of experience by noise level, we can observe the nearly equal spacing between the three lines, thus confirming our conclusions from the AOV table. The profile plots depicting the interactions of gender and experience and of gender and noise level again confirm the tests from the AOV table. The lines are no longer equally spaced. The difference in the mean performance indices between females and males increases with increasing experience. The difference in the mean performance indices between females and males is relatively large for low levels of experience, but male and female performance is nearly equal at the higher level of experience.

FIGURE 14.10(a)
Profile plot of gender by noise level

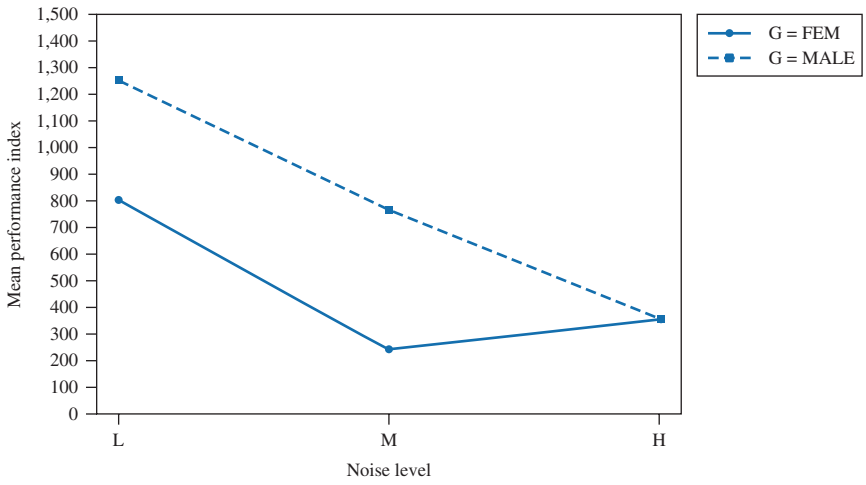


FIGURE 14.10(b)
Profile plot of gender by experience

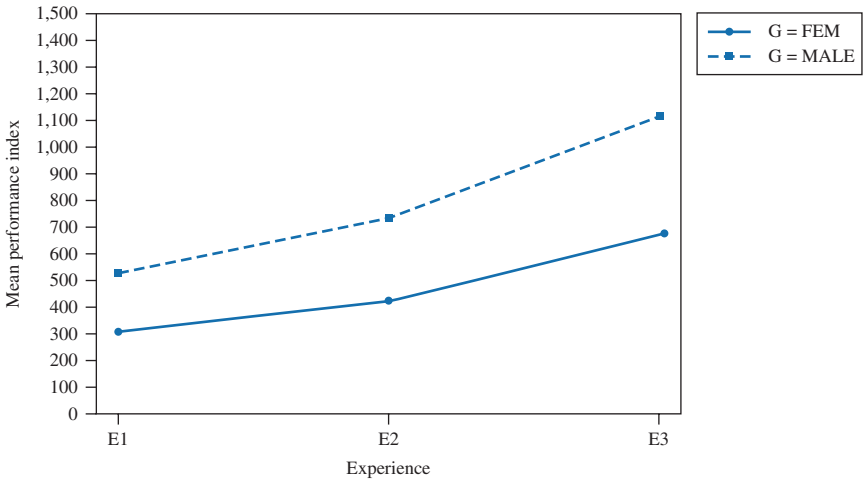
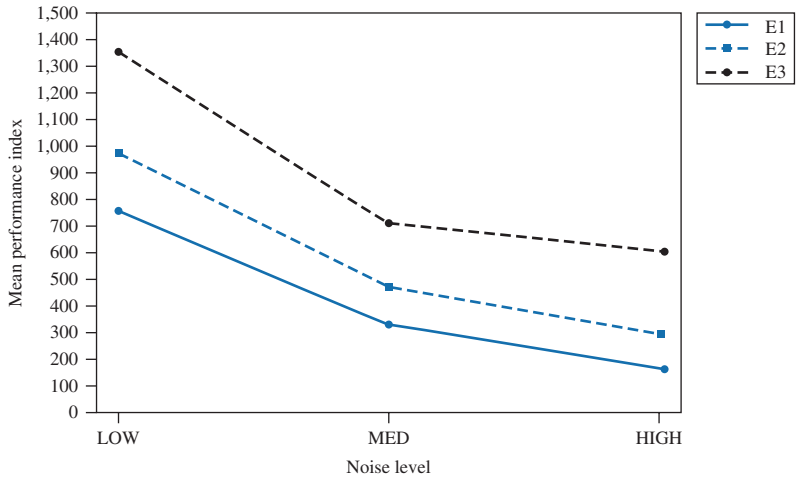


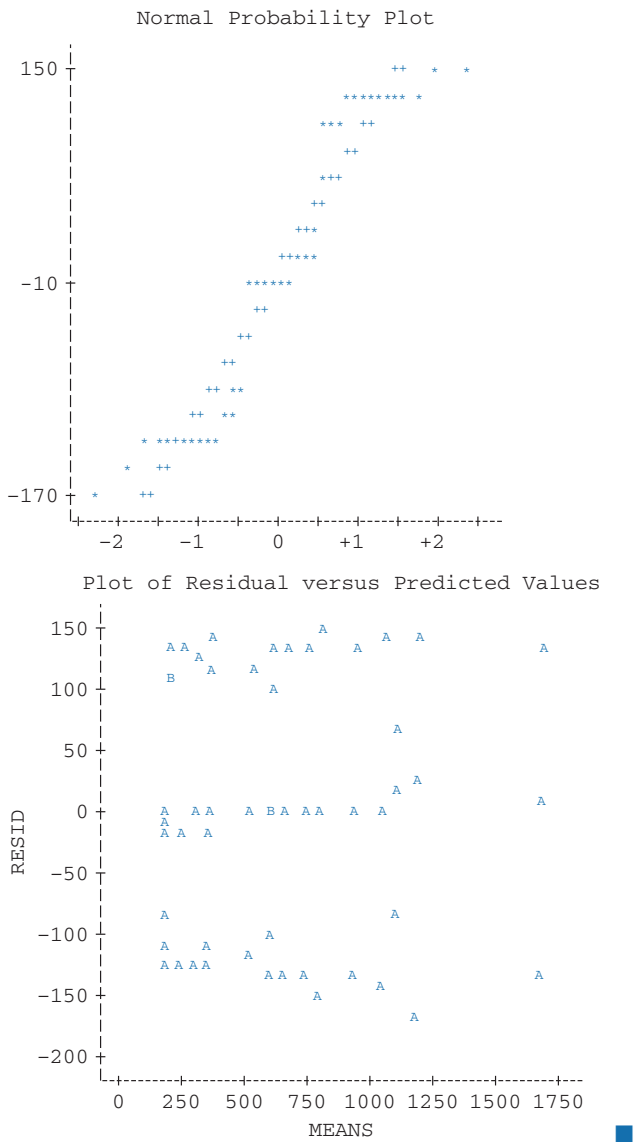
FIGURE 14.10(c)
Profile plot of noise level by experience



The output contains summary statistics, a plot of the residuals versus the predicted value, and a normal probability plot of the residuals. Although the tests of the normality of the residuals appear to indicate nonnormality, the plots do not indicate a strong deviation from a normal distribution. The residual plots do not indicate a violation of the equal variance condition, as the spread in the residuals appears nearly constant with increasing values of the predicted performance index.

```
Variable: RESID

N                          54    Sum Weights                54
Mean                        0    Sum Observations            0
Std Deviation     103.834714    Variance           10781.6478
Skewness          0.00981289    Kurtosis           -1.4081028

                    Tests for Normality

    Test              --Statistic---  -----p Value-----

    Shapiro-Wilk       W    0.885132  Pr < W    <0.0001
    Anderson-Darling   A-Sq 2.150291  Pr > A-Sq <0.0050
```

```
Stem Leaf                #  Boxplot
  14 026                 3     |
  12 711355778           9     |
  10 17948               5  +------+
   8                          |      |
   6 1                    1  |      |
   4                          |      |
   2 4                    1  |      |
   0 11453                5  |   +  |
  -0 944443222111        12  *------*
  -2                         |      |
  -4                         |      |
  -6                         |      |
  -8 73                  2   |      |
 -10 8270                4   +------+
 -12 9755306541         10      |
 -14 0                   1      |
 -16 3                   1      |
    ----+----+----+----+
Multiply Stem.Leaf by 10**+1
```



14.4 Factorial Treatment Structures with an Unequal Number of Replications

The analysis of a completely randomized design with an unequal number of replications for the $t = ab$ factorial treatments is more complex than the analysis with equal number of replications. Suppose we have a two-factor experiment with factor A having a levels and factor B having b levels. Let n_{ij} be the number of replications for the treatment consisting of the ith level of factor A and the jth level of factor B. Generally, the n_{ij}s are designed to be the same value for all treatments; that is, $n_{ij} = r$ for $i = 1, \ldots, a$ and $j = 1, \ldots, b$. There may be special reasons to design an experiment having unequal replications; for example, added information is required for certain combinations of the factor levels. However, in most cases, the unequal number of replications occurs due to problems that arise during the implementation of the experiment. Laboratory animals die, animals jump fences and destroy the crops on selected plots, volunteers decide not to participate in a study, or there is an unequal response rate in a study involving a mailed questionnaire. In all these situations, the researcher ends up with a data set having an unequal number of replications. This results in several problems. The formulas for the sums of squares for main effects and interactions are no longer valid. The estimation of the marginal means, $\mu_{i.}$ and $\mu_{.j}$, are no longer just the corresponding sample means. The sum of squares for the main effects of factors A and B added to the sum of squares for interaction no longer total the model sum of squares. This is due to the nonorthogonality of the contrasts that compose these sum of squares. In these situations, we must rely on computer software to produce the AOV tables and the estimated main effects and their standard errors.

In a completely randomized design with a single factor having t levels and n_i replications, the treatment means are estimated by

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_{i\cdot}.$$

The estimated standard errors of the estimated treatment means are given by $\widehat{\mathrm{SE}}(\hat{\mu}_i) = \sqrt{\mathrm{MSE}/n_i}$.

The tests of hypotheses and estimators are similar for designs with equal or unequal numbers of replications, provided $n_i > 1$ for all $i = 1, \ldots, t$. The only difference is that the treatments with larger numbers of replications will have a more precise estimate of their mean and a smaller estimated standard error. The testing procedures are similar for equally and unequally replicated experiments.

When we have designs with factorial treatments, the test statistics and estimation of marginal treatment means differ depending on whether we have equal or unequal numbers of replications. With equal replications, we can use the formulas given in Section 14.3 to obtain the sums of squares for main effects and interactions. When the experiment involves unequal replications, it is necessary to use computer software to obtain those sums of squares.

The estimation of treatment means pose a similar problem. When we have equal replications, the estimates of the treatment means and marginal means are the corresponding sample means. However, in the case of unequal replications, this is no longer true. We will illustrate these formulas for the case of a two-factor experiment with factor A having a levels, factor B having b levels, and the number of replications, n_{ij}, depending on the particular factor-level combinations. The sample estimates of the treatment means, μ_{ij}, are the same as in the equal replications

case, but the estimates of the marginal means, $\mu_{i.}$ and $\mu_{.j}$, are different from the equal replications case.

The least-squares estimates are given here:

1. Treatment mean, μ_{ij}: $\hat{\mu}_{ij} = \bar{y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}$:

   The formula is the same for the equal replications case: $n_{ij} = r$

2. Factor A marginal mean, $\mu_{i.} = \frac{1}{b}\sum_{j=1}^{b} \mu_{ij}$:

$$\hat{\mu}_{i.} = \frac{1}{b}\sum_{j=1}^{b} \hat{\mu}_{ij} = \frac{1}{b}\sum_{j=1}^{b} \bar{y}_{ij.} = \frac{1}{b}\sum_{j=1}^{b} \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}$$

3. Factor B marginal mean, $\mu_{.j} = \frac{1}{a}\sum_{i=1}^{a} \mu_{ij}$:

$$\hat{\mu}_{.j} = \frac{1}{a} \sum_{i=1}^{a} \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}$$

From the above formula, we can see that when $n_{ij} = r$ for all (i, j):

$$\hat{\mu}_{i.} = \frac{1}{b}\sum_{j=1}^{b} \frac{1}{r}\sum_{k=1}^{r} y_{ijk} = \frac{1}{br}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk} = \bar{y}_{i..}$$

Similarly, $\hat{\mu}_{.j} = \bar{y}_{.j.}$ when $n_{ij} = r$ for all (i, j). Thus, care must be taken when dealing with factorial treatment structures with unequal replications. We will illustrate these ideas using the following example.

EXAMPLE 14.8

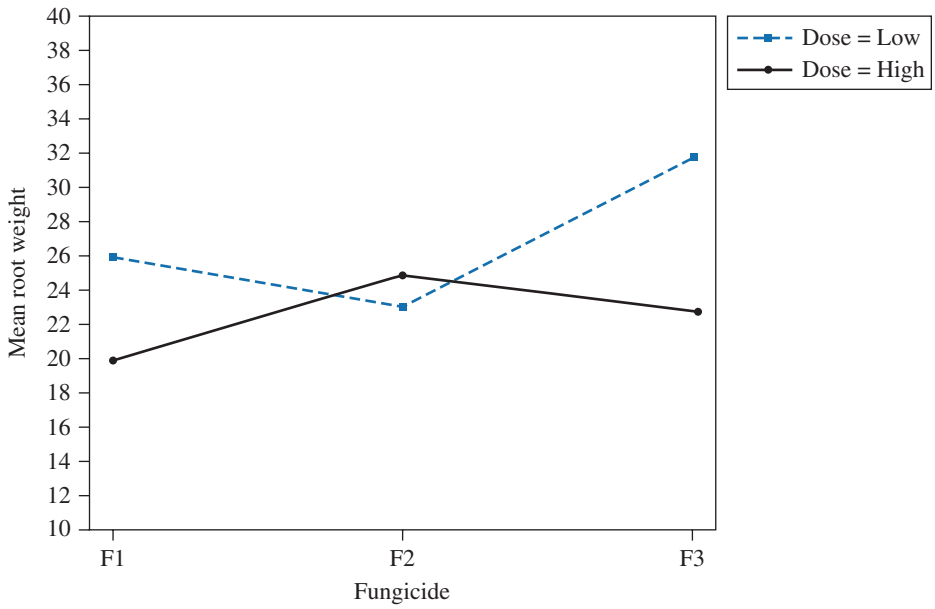
A horticulturist is interested in studying the effectiveness of fungicide treatments applied to plots on which roses are grown. Six treatments, consisting of one of three types of fungicide at one of two dose levels, were randomly assigned to 24 plots. This is a completely randomized design with a factorial (2×3) treatment structure and $r = 4$ replications per treatment. Rose plants of the same health, size, and age were inoculated, planted, and, after 20 weeks, dug up and the root weights determined. However, a number of plants died during the 20 weeks. This resulted in an unbalanced design with the number of replications per treatment varying from $n_{ij} = 2$ to $n_{ij} = 4$ (see Table 14.22).

TABLE 14.22
Root weight data

	Fungicide		
Dose Level	**1**	**2**	**3**
1	19	24	22
	20	26	25
	21		25
			19
2	25	21	31
	27	24	32
		24	33
			32

A profile plot is given in Figure 14.11. There appears to be an interaction between the two factors in that the two lines intersect. However, we need to test if there is significant evidence of an interaction after taking into account the level of variation in the estimation of the treatment means.

FIGURE 14.11
Profile plot of fungicide treatments



The test of hypotheses and estimation of the treatment means will be obtained from the following output from SAS:

```
Class          Levels      Values
A                   2      1 2
B                   3      1 2 3

Dependent Variable: Y
                                      Sum of
Source                     DF        Squares    Mean Square    F Value    Pr > F
Model                       5    305.2500000     61.0500000      18.91    <.0001
Error                      12     38.7500000      3.2291667
Corrected Total            17    344.0000000

Source                     DF    Type III SS    Mean Square    F Value    PR > F
A                           1    81.02884615    81.02884615      25.09    0.0003
B                           2    67.92272727    33.96136364      10.52    0.0023
A*B                         2    95.74090909    47.87045455      14.82    0.0006

The GLM Procedure                                          Least Squares Means

A            N            Mean         Std Dev        LSMEAN      Standard Error
1            9      22.3333333      2.73861279      22.5833333      0.6234549
2            9      27.6666667      4.41588043      27.0000000      0.6234549

B            N            Mean         Std Dev        LSMEAN      Standard Error
1            5      22.4000000      3.43511281      23.0000000      0.8202092
2            5      23.8000000      1.78885438      24.0000000      0.8202092
3            8      27.3750000      5.31675250      27.3750000      0.6353313

A   B        N            Mean         Std Dev        LSMEAN      Standard Error
1   1        3      20.0000000      1.00000000      20.0000000      1.0374916
1   2        2      25.0000000      1.41421356      25.0000000      1.2706626
1   3        4      22.7500000      2.87228132      22.7500000      0.8984941
2   1        2      26.0000000      1.41421356      26.0000000      1.2706626
2   2        3      23.0000000      1.73205081      23.0000000      1.0374916
2   3        4      32.0000000      0.81649658      32.0000000      0.8984941
```

From the SAS output, we obtain p-value = .0006 for testing for an interaction between factors A and B. This confirms our observations from the profile plot.

Using our formulas for a balanced design,

$$\text{SSA} = \sum_{i=1}^{a} n_{i.}(\bar{y}_{i..} - \bar{y}_{...})^2 \qquad \text{SSB} = \sum_{i=1}^{b} n_{.j}(\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$\text{SSAB} = \sum_{i=1}^{a}\sum_{i=1}^{b} n_{ij}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

we would obtain the following values for SSA, SSB, and SSAB:

$$\text{SSA} = 128 \qquad \text{SSB} = 86.125 \qquad \text{SSAB} = 98.5917$$

These values certainly do not agree with the values given in the previous AOV table. The reason for the disagreement is that the least-squares estimates of the marginal means are not equal to the corresponding sample means. We will demonstrate this result using the sample means in Table 14.23.

TABLE 14.23
Treatment sample means

	Fungicide			
Dose Level	**1**	**2**	**3**	$\bar{y}_{i..}$
1	20	25	22.75	22.333
2	26	23	32	27.667
$\bar{y}_{.j.}$	22.4	23.8	27.375	

The least-squares estimates of the treatment means, $\hat{\mu}_{ij}$, are equal to $\bar{y}_{ij.}$ for all six treatments. However, the least-squares estimates of the treatment marginal means, $\mu_{i.}$ and $\mu_{.j}$, are given by

$$\hat{\mu}_{1.} = \frac{1}{3}\sum_{j=1}^{3} \hat{\mu}_{1j} = \frac{1}{3}[20 + 25 + 22.75] = 22.583 \neq 22.333 = \bar{y}_{1..}$$

$$\hat{\mu}_{2.} = \frac{1}{3}\sum_{j=1}^{3} \hat{\mu}_{2j} = \frac{1}{3}[26 + 23 + 32] = 27 \neq 27.667 = \bar{y}_{2..}$$

$$\hat{\mu}_{.1} = \frac{1}{2}\sum_{i=1}^{2} \hat{\mu}_{i1} = \frac{1}{2}[20 + 26] = 23 \neq 22.4 = \bar{y}_{.1.}$$

$$\hat{\mu}_{.2} = \frac{1}{2}\sum_{i=1}^{2} \hat{\mu}_{i2} = \frac{1}{2}[25 + 23] = 24 \neq 23.8 = \bar{y}_{.2.}$$

$$\hat{\mu}_{.3} = \frac{1}{2}\sum_{i=1}^{2} \hat{\mu}_{i3} = \frac{1}{2}[22.75 + 32] = 27.375 = 27.375 = \bar{y}_{.3.}$$

In general, the least-squares estimates of the treatment marginal means are not equal to the corresponding sample means, $\hat{\mu}_{i.} \neq \bar{y}_{i..}$ and $\hat{\mu}_{.j} \neq \bar{y}_{.j.}$, although occasionally the two estimates will agree, as is seen for $\hat{\mu}_{.3}$. ■

When all of the data for some treatments are completely deleted or missing in an experiment—that is, $n_{ij} = 0$ for some combinations (i, j)—the standard analysis of the experiment will often lead to very misleading conclusions. The AOV table in the output from most software packages will provide sums of squares and tests that are not very meaningful. An excellent reference for the analysis of this type of experiment is the book *Analysis of Messy Data* (Milliken and Johnson, 2009). Consider the following example from this book.

EXAMPLE 14.9

A bakery scientist wanted to study the effects of combining three different fats (F_1) with each of three surfactants (F_2) on the specific volume of bread loaves baked from doughs mixed from each of the nine treatment combinations. Four loaves were made from each of the nine treatment combinations. Unfortunately, one container of yeast turned out to be ineffective, and the data from the 15 loaves made with that yeast had to be removed from the analysis. The data are given in Table 14.24.

TABLE 14.24
Specific volumes from baking experiment

	Factors		Loaf					
Treatment	Fat	Surfactant	1	2	3	4	n_{ij}	$\bar{y}_{ij.}$
1	1	1	6.7	4.3	5.7	*	3	5.57
2	1	2	7.1	*	5.9	5.6	3	6.20
3	1	3	*	*	*	*	0	*
4	2	1	*	5.9	7.4	7.1	3	6.80
5	2	2	*	*	*	*	0	*
6	2	3	6.4	5.1	6.2	6.3	4	6.00
7	3	1	7.1	5.9	*	*	2	6.50
8	3	2	7.3	6.6	8.1	6.8	4	7.00
9	3	3	*	7.5	9.1	*	2	8.30
Total							21	6.58

This experiment is a completely randomized design with a 3×3 factorial treatment structure and four replications. However, a number of the replications are not observed. This results in several treatments having no observations in the experiment. Often experimental data such as those in Table 14.24 are analyzed using computer software. The following analysis from SAS demonstrates the problems that result from such an analysis.

The model used in the following analysis is given here:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + e_{ijk} \quad \text{with} \quad i = 1, 2, 3;\ j = 1, 2, 3$$

This was designed as an equally replicated experiment with $r = 4$; however, because of problems that arose during the experiment, the numbers of actual observations per treatment are given below:

$$n_{11} = 3; n_{12} = 3; n_{13} = 0; n_{21} = 3; n_{22} = 0; n_{23} = 4; n_{31} = 2; n_{32} = 4; n_{33} = 2$$

The following output obtained from SAS contained no specification of missing treatments.

```
                    Analysis as a CR 3x3 factorial

Class        Levels        Values
fat               3        1 2 3
surf              3        1 2 3
Number of observations       36

NOTE: Due to missing values, only 21 observations can be used in this
      analysis.

Dependent Variable: sv

                                    Sum of
Source                      DF      Squares     Mean Square   F Value  Pr > F
Model                        6   12.47142857      2.07857143      2.95  0.0447
Error                       14    9.86666667      0.70476190
Corrected Total             20   22.33809524
```

```
Source                     DF       Type I SS     Mean Square   F Value   Pr > F
fat                         2     7.45261905      3.72630952      5.29   0.0195
surf                        2     0.29722997      0.14861498      0.21   0.8124
fat*surf                    2     4.72157956      2.36078978      3.35   0.0647

Source                     DF      Type II SS     Mean Square   F Value   Pr > F
fat                         2     6.47812282      3.23906141      4.60   0.0292
surf                        2     0.29722997      0.14861498      0.21   0.8124
fat*surf                    2     4.72157956      2.36078978      3.35   0.0647

Source                     DF     Type III SS     Mean Square   F Value   Pr > F
fat                         2     6.00174091      3.00087046      4.26   0.0359
surf                        2     0.99963357      0.49981678      0.71   0.5089
fat*surf                    2     4.72157956      2.36078978      3.35   0.0647

Source                     DF      Type IV SS     Mean Square   F Value   Pr > F
fat                        2*     3.87252033      1.93626016      2.75   0.0985
surf                       2*     1.67022222      0.83511111      1.18   0.3346
fat*surf                    2     4.72157956      2.36078978      3.35   0.0647

* NOTE: Other Type IV Testable Hypotheses exist which may yield
        different SS.

     Least Squares Means
                                  Standard
     fat       sv LSMEAN             Error    Pr > |t|
     1           Non-est         .               .
     2           Non-est         .               .
     3        7.33333333      0.31286355        <.0001

                                  Standard
     surf      sv LSMEAN             Error    Pr > |t|
     1        6.28888889      0.30225490        <.0001
     2           Non-est         .               .
     3           Non-est         .               .


                                        Standard                    LSMEAN
     fat    surf       sv LSMEAN           Error     Pr > |t|      Number
     1      1         5.56666667      0.48468612       <.0001           1
     1      2         6.20000000      0.48468612       <.0001           2
     2      1         6.80000000      0.48468612       <.0001           3
     2      3         6.00000000      0.41975049       <.0001           4
     3      1         6.50000000      0.59361684       <.0001           5
     3      2         7.20000000      0.41975049       <.0001           6
     3      3         8.30000000      0.59361684       <.0001           7
```

Type III and IV sums of squares are the mostly widely used in the analysis of experiments. They test the type of hypotheses of most interest to experimenters. When some of the treatments are not observed in the experiment—that is, $n_{ij} = 0$ for some treatments—the Type IV sum of squares adjusts factor effects by averaging over one or more common levels of the other factor effects. In most cases, when some treatments are not observed, the Type IV sum of squares is testing hypotheses that are most likely to have reasonable interpretations. However, as is true for all four types of sums of squares, it is difficult to determine the actual hypotheses being tested. There are many other possible Type IV hypotheses that can be generated. PROC GLM in SAS automatically generates a set of Type IV hypotheses. Thus, it is impossible to interpret the significance of the effects using the p-value for the main and interaction effects because the set of hypotheses tested is not displayed. The interpretation problem is shown in the SAS output with the display of the statement "Other Type IV Testable Hypotheses exist which may yield different SS."

The more appropriate methodology is to ignore the factorial structure of the treatments and just consider the experiment as having a single factor with t levels. For example, in Example 14.9, the original design had $t = (3)(3)$ treatments. However, after the completion of the experiment, only seven of the nine treatments were observed. Thus, we should analyze the data from the experiment as if there was just a single factor having $t = 7$ treatments. It is still possible in many such experiments to construct contrasts that are testing hypotheses that are directly of interest to the researcher. Consider the following analysis.

Let y_{ijk} = specific volume of the kth loaf using ith level of fat and jth level of surfacant.

Model: $y_{ijk} = \mu_{ij} + e_{ijk}$; for $i, j = 1, 2, 3$; $k = 1, \ldots, r_{ij}$

$$\text{SS}_{\text{TOT}} = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{r_{ij}}[y_{ijk} - \bar{y}_{\ldots}]^2 = 22.338095 \quad \text{df}_{\text{TOT}} = 21 - 1 = 20$$

$$\text{SS}_E = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{r_{ij}}[y_{ijk} - \hat{\mu}_{ij}]^2 = 9.8666667 \quad \text{df}_E = N - t = 21 - 7 = 14$$

$$\text{SS}_{\text{MODEL}} = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{r_{ij}}[\hat{\mu}_{ij} - \bar{y}_{\ldots}]^2 = 12.471429 \quad \text{df}_M = t - 1 = 7 - 1 = 6$$

We want to decompose SS_{MODEL} into terms that represent differences in the $t = 7$ treatments: fat (F) main effect, surfactant (S) main effect, and F × S interaction.

I. First, test for overall difference in the seven treatments:

Test H_0: $\mu_{11} = \mu_{12} = \mu_{21} = \mu_{23} = \mu_{31} = \mu_{32} = \mu_{33}$ versus H_a: Not all μ_{ij} are equal.

$$F = \frac{\text{MS}_{\text{MODEL}}}{\text{MS}_E} = \frac{12.471429/6}{9.866667/14} = 2.95 \text{ with df} = 6, 14 \Rightarrow p\text{-value} = .0447$$

Therefore, there appears to be some evidence of a difference in the seven treatment means.

II. Construct contrasts that represent comparisons between treatment means that are main effects and two-way interactions:

Table 14.25 contains eight mutually orthogonal contrasts that would represent the $t - 1 = 9 - 1 = 8$ df for decomposing SS_{MODEL} into components for main effects and interaction provided all nine treatments were observed.

Because not all factor combinations were observed, the contrasts which represent main effects and interactions are modified to the contrasts given in Table 14.26.

The choices for the contrasts are not unique as is illustrated with three possible sets of contrasts for evaluating the main effect of surfacant. Furthermore, the set of six contrasts is not a set of orthogonal contrasts.

The determination of whether there is significant evidence of a main effect for fat or surfacant and whether there is significant evidence of an interaction between fat and surfacant relies on testing the significance of the contrasts in Table 14.26. The following SAS output contains the tests for the six contrasts displayed in Table 14.26.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Main Fat	2	3.87252033	1.93626016	2.75	0.0985
Main Surf	2	1.67022222	0.83511111	1.18	0.3346
Interaction	2	4.72157956	2.36078978	3.35	0.0647

TABLE 14.25
Coefficients for mutually orthogonal contrasts in nine treatment means

Contrast	Effect	Treatment Means								
		μ_{11}	μ_{12}	μ^*_{13}	μ_{21}	μ^*_{22}	μ_{23}	μ_{31}	μ_{32}	μ_{33}
Main fat	C_1	1	1	1	−1	−1	−1	0	0	0
	C_2	1	1	1	1	1	1	−2	−2	−2
Main surf.	C_3	1	−1	0	1	−1	0	1	−1	0
	C_4	1	1	−2	1	1	−2	1	1	−2
Interaction	C_5	1	−1	0	−1	1	0	0	0	0
	C_6	1	1	−2	−1	−1	2	0	0	0
	C_7	1	−1	0	1	−1	0	−2	2	0
	C_8	1	1	−2	1	1	−2	−2	−2	4

Note: * indicates that treatment was not observed.

TABLE 14.26
Coefficients for contrasts in observed seven treatment means

Contrast	Effect	Treatment Means						
		μ_{11}	μ_{12}	μ_{21}	μ_{23}	μ_{31}	μ_{32}	μ_{33}
Main, fat	C_1	1	1	0	0	−1	−1	0
	C_2	0	0	1	1	−1	0	−1
Main, surf. 1	C_3	1	−1	0	0	1	−1	0
	C_4	0	0	1	−1	1	0	−1
Main, surf. 2	C_3	0	0	0	0	1	0	−1
	C_4	0	0	0	0	1	−2	1
Main, surf. 3	C_3	0	0	0	0	0	1	−1
	C_4	0	0	1	−1	1	0	−1
Interaction	C_5	1	−1	0	0	−1	1	0
	C_6	0	0	1	−1	−1	0	1

From the previous output, we can observe that there is not significant evidence of pseudo-main effects and pseudo-interaction in this experiment. The *p*-values for the six contrasts in the SAS output are identical to the *p*-values associated with the Type IV sum of squares from the AOV table in the SAS output. Thus, we would reach the same conclusions that we reached using the SAS output. The important point is that using the contrast approach, we know what hypotheses are being tested, whereas the exact hypotheses being tested by the Type IV sum of squares may vary from analysis to analysis. Furthermore, the output from other software packages may not produce the Type IV sum of squares produced by SAS, so the researcher would not know the hypotheses being tested using the AOV *F* tests. Thus, no matter what software package is used to analyze the data, there is not direct information concerning what hypotheses are being tested when some of the factorial combinations are not observed in the experiment.

14.5 Estimation of Treatment Differences and Comparisons of Treatment Means

We have emphasized the analysis of variance associated with factorial experiments. However, there are times when we might be more interested in estimating the difference in mean responses for two treatments (different levels of the same factor or different combinations of levels). For example, an environmental engineer might be more interested in estimating the difference in the mean dissolved oxygen contents for a lake before and after rehabilitative work than in testing to see

whether there is a difference. Thus, the engineer is asking the question "What is the difference in mean dissolved oxygen contents?" instead of the question "Is there a difference between the mean contents before and after the cleanup project?"

The Tukey procedure can be used to evaluate the difference in treatment means for a k-factor treatment structure in a completely randomized design. Let $\bar{y}_i$ denote the mean response for treatment i, $\bar{y}_{i'}$ denote the mean response for treatment i', and n_t denote the number of observations in each treatment. A set of simultaneous $100(1-\alpha)\%$ confidence intervals on $\mu_i - \mu_{i'}$, the difference in mean responses for the two treatments, is defined as shown here.

$100(1-\alpha)\%$ Confidence Interval for the Difference in t Treatment Means

$$(\bar{y}_i - \bar{y}_{i'}) \pm q_\alpha(t, \nu)\sqrt{\frac{s_\varepsilon^2}{n_t}}$$

where s_ε^2 is the square root of MSE in the AOV table and $q_\alpha(t, \nu)$ can be obtained from Table 10 in the Appendix for the specified α and ν, the degrees of freedom for MSE.

EXAMPLE 14.10

A company was interested in comparing three different display panels for use by air traffic controllers. Each display panel was to be examined under five different simulated emergency conditions. Thirty highly trained air traffic controllers with similar work experience were enlisted for the study. A random assignment of controllers to display panel–emergency conditions was made, with two controllers assigned to each factor–level combination. The time (in seconds) required to stabilize the emergency situation was recorded for each controller. These data appear in Table 14.27.

TABLE 14.27 Display panel data (time in seconds)

	Emergency Condition, A				
Display Panel, B	**1**	**2**	**3**	**4**	**5**
1	18.8	32.7	25.1	41.7	14.9
	15.2	33.3	23.9	33.3	12.1
2	14.6	36.5	23.9	38.0	14.7
	13.4	26.5	21.1	35.0	11.3
3	27.8	45.0	40.8	55.0	29.4
	24.2	43.0	36.2	54.0	22.6

a. Construct a profile plot.
b. Run an analysis of variance that includes a test for interaction.

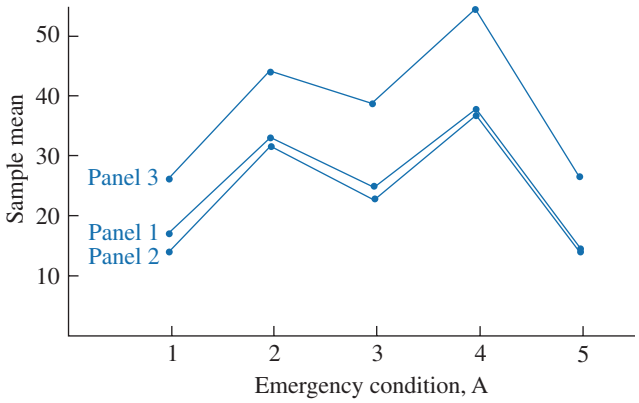
Solution

a. The sample means are given in Table 14.28 and then displayed in a profile plot in Figure 14.12. From the profile plot, we observe that

TABLE 14.28 Mean reaction times for display panel–emergency condition study

	Emergency Condition, A					
Display Panel, B	**1**	**2**	**3**	**4**	**5**	**Means $\bar{y}_{j.}$**
1	17	33	24.5	37.5	13.5	25.1
2	14	31.5	22.5	36.5	13	23.5
3	26	44	38.5	54.5	26	37.8
Means $\bar{y}_{i..}$	19.0	36.2	28.5	42.8	17.5	$\bar{y}... = 28.8$

FIGURE 14.12
Plot of panel means for each emergency condition



the difference in mean reaction times for controllers on any pair of different display panels remains relatively constant across all five emergency conditions. Panel 1 and panel 2 yield essentially the same mean reaction times across the five emergency conditions, whereas panel 3 produces mean reaction times that are consistently higher than the mean times for the other two panels. We will next confirm these observations using tests of hypotheses that take into account the variability of the reaction times about the observed mean times.

b. The computer output for the analysis of variance table is given in Table 14.29.

TABLE 14.29
AOV table for display panel–emergency condition study

General Linear Models Procedure

Dependent Variable: y, Stabilization Time

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	4122.8000	294.4857	28.65	0.0001
Error	15	154.1800	10.2787		
Corrected Total	29	4276.9800			

R-Square	C.V.	Root MSE	Y Mean
0.963951	11.132087	3.2060	28.800

Source	DF	Type I SS	Mean Square	F Value	Pr > F
D	2	1227.8000	613.9000	59.73	0.0001
E	4	2850.1333	712.5333	69.32	0.0001
D*E	8	44.8667	5.6083	0.55	0.8049

The first test of hypotheses is for an interaction between the two factors, emergency condition and type of display panel. The computed value of $F = .55$ is less than the critical value of F, 2.64, for $\alpha = .05$, $df_1 = 8$, and $df_2 = 15$. Thus, have insufficient evidence (p-value = .8049) to indicate an interaction between emergency conditions and type of display panel. This confirms our observations from the profile plot. Because the interaction was not significant, we will next test for a main effect due to type of display panel. The computed value of F, 59.73, is more than the critical value of F, 3.68, for $\alpha = .05$, $df_1 = 2$, and $df_2 = 15$, so we have sufficient evidence (p-value $< .0001$) to indicate a significant difference in mean reaction times across the three types of display panels. ■

EXAMPLE 14.11

Refer to Example 14.10. The researchers were very interested in the size of the differences in mean reaction times among the three types of panels. Estimate these differences using 95% confidence intervals.

Solution Because there is not a significant interaction between type of display panel and type of emergency condition, the sizes of the differences in mean reaction times among the types of display panels would be relatively the same for all five types of emergency conditions. Thus, we can examine the main effect means for the three display panels, averaging over the five emergency conditions: $\hat{\mu}_{\cdot j} = \bar{y}_{\cdot j\cdot}$, for $j = 1, 2, 3$. From Table 14.28, we have

$$\bar{y}_{\cdot 1\cdot} = 25.1 \quad \bar{y}_{\cdot 2\cdot} = 23.5 \quad \bar{y}_{\cdot 3\cdot} = 37.8$$

The value of $q_\alpha(t, \nu)$ for $\alpha = .05$, $t = 3$, and $\nu = 15$ is 3.67; the estimate of σ_ε is

$$s_\varepsilon = \sqrt{\text{MSE}} = \sqrt{10.2787} = 3.21$$

The formula for a 95% confidence interval on the difference between the mean reaction times of two display panels, $\mu_{\cdot j} - \mu_{\cdot j'}$, is given by

$$\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot j'\cdot} \pm q_{.05}(3, 15)\sqrt{\frac{s_\varepsilon^2}{n_t}}$$

For panels 2 and 3, we have $n_t = 10$ observations per panel; thus, we have

$$37.8 - 23.5 \pm 3.67\sqrt{\frac{10.2787}{10}}$$

$$14.3 \pm 3.72$$

that is, 10.58 to 18.02. Therefore, we are 95% confident that the difference in the mean reaction times between display panel 2 and display panel 3 is between 10.58 and 18.02 seconds. Similarly, we can calculate confidence intervals on the differences between panels 1 and 3 and between panels 1 and 2. ■

After determining that there was a significant main effect using the F test, we would proceed with two further inference procedures. First, we would place confidence intervals on the difference between any pair of factor–level means—$\mu_{i\cdot} - \mu_{i'\cdot}$ for factor A or $\mu_{\cdot j} - \mu_{\cdot j'}$ for factor B—using the procedure illustrated in Example 14.11. This would estimate the effect sizes for these two factors. Next, we would want to determine which pairs of factor-level means are significantly different.

As discussed in Chapter 9, we would apply one of the **multiple-comparison procedures** in order to control the experimentwise error rate for comparing the pairs of factor levels. There would be $a(a-1)/2$ pairs for factor A and $b(b-1)/2$ pairs for factor B. The choice of which procedure to use would once again depend on the experiment, as discussed in Chapter 9. All of the procedures discussed in Chapter 9, such as Tukey Scheffé, and Bonferroni, can be performed for a k-factor treatment structure in a completely randomized experiment. The quantity s_w^2 in the formulas given in Chapter 9 for these procedures is replaced with MSE, the degrees for MSE are obtained from the AOV table, and the sample size n refers to the number of observations per mean value in the comparison—that is, the number of data values averaged to obtain $\bar{y}_{i\cdot\cdot}$, for example.

EXAMPLE 14.12

Refer to Example 14.10 and the data in Tables 14.27 and 14.28. Use Tukey's W procedure to locate significant differences among display panels.

Solution For the Tukey's W procedure, we use the formula presented in Chapter 9:

$$W = q_\alpha(t, \nu)\sqrt{\frac{s_w^2}{n}}$$

where s_w^2 is MSE from the AOV table, based on $\nu = 15$ degrees of freedom, and $q_\alpha(t, \nu)$ is the upper-tail critical value of the studentized range for comparing t different population means. The value of $q_\alpha(t, \nu)$ from Table 10 in the Appendix for comparing the three display panel means, each of which has 10 observations per sample mean, is

$$q_{.05}(3, 15) = 3.67$$

For 10 observations per mean, the value of W is

$$W = q_\alpha(t, \nu)\sqrt{\frac{s_w^2}{n}} = 3.67\sqrt{\frac{10.28}{10}} = 3.72$$

The display panel means are, from Table 14.28,

$$\bar{y}_{.1.} = 25.1 \quad \bar{y}_{.2.} = 23.5 \quad \bar{y}_{.3.} = 37.8$$

First, we rank the sample means from lowest to highest:

Display panel	2	1	3
Means	23.5	25.1	37.8

For the two means that differ (in absolute value) by more than $W = 3.72$, we declare them to be significantly different from each other. The results of our multiple-comparison procedure are summarized here:

Display panel	2	1	3

Thus, display panels 1 and 2 both have mean reaction times significantly lower than display panel 3, but we are unable to detect a difference in the mean reaction times between panels 1 and 2. ■

14.6 Determining the Number of Replications

The number of replications in an experiment is the crucial element in determining the accuracy of estimators of the treatment means and the power of tests of hypotheses concerning differences between the treatment means. In most situations, the greater the number of replications, the greater the accuracy of the estimators, the more precise the confidence intervals on treatment means, and the greater the power of the tests of hypotheses. The conditions that constrain the researcher from using very large numbers of replications are the cost of running the experiment, the time needed to handle a large number of experimental units, and the availability of experimental units. Thus, the researcher must determine the minimum number of replications required to meet reasonable specifications on the accuracy of estimators or on the power of tests of hypotheses.

Using the Accuracy of Estimator Specifications to Determine the Number of Replications

We can determine the number of replications by specifying the desired width of a $100(1-\alpha)\%$ confidence interval on the treatment mean. In Chapter 5, we provided a formula for determining the sample size needed so that we were $100(1-\alpha)\%$ confident that the sample estimate was within E units of the true treatment mean. If we let r be the number of replications, σ be the experimental standard deviation, and E be the desired accuracy of the estimator, then we can approximate the value of r using the following formula.

Sample Size r Required to Be $100(1-\alpha)\%$ Confident That the Estimator Is Within E Units of the Treatment Mean μ

$$r = \frac{(z_{\alpha/2})^2 \hat{\sigma}^2}{E^2}$$

In using this formula, the experimenter must specify

1. The desired level of confidence, $100(1-\alpha)\%$.
2. The level of precision, E.
3. An estimate of σ. The estimate of σ may be obtained from a pilot study, similar past experiments, or literature on similar experiments, or a rough estimator can be used: $\hat{\sigma} = (\text{largest value} - \text{smallest value})/4$. The following example will illustrate these calculations.

EXAMPLE 14.13

A researcher is designing a project to study the yield of pecans under four rates of nitrogen application. The researcher wants to obtain estimates of the treatment means μ_1, μ_2, μ_3, and μ_4 such that she will be 95% confident that the estimates are within 4 pounds of the true mean yield. She wants to determine the necessary number of replications to achieve these goals.

Solution From previous experiments, the yields have ranged from 40 pounds to 70 pounds. Thus, an estimate of σ is given by

$$\hat{\sigma} = \frac{70 - 40}{4} = 7.5$$

From the normal tables, $z_{.025} = 1.96$. The value of E is 4 pounds, as specified by the researcher. Thus, we determine that the number of replications is

$$r = \frac{(z_{\alpha/2})^2 \hat{\sigma}^2}{E^2} = \frac{(1.96)^2 (7.5)^2}{(4)^2} = 13.51$$

Thus, the researcher should use 14 replications on each of the treatments to obtain the desired precision. ■

Using this technique to determine the number of replications does not take into account the power of the F test to detect specified differences in the treatment means. Thus, the following method of determining the number of replications is preferred in most studies.

Using the Power of the *F* Test to Determine the Number of Replications

In a study involving t treatments, one of the goals is to test the hypotheses

H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$

H_a: Not all μs are equal.

The test procedure is to reject H_0 if $F \geq F_{\alpha,\, t-1,\, N-t}$, with $F = \text{MST/MSE}$, where MST and MSE are the mean squares from the AOV table. The number of replications, with $r_1 = r_2 = \cdots = r_t = r$, will be determined by specifying the following parameters with respect to the test statistic:

1. The significance level, α
2. The size of the difference $D = |\mu_i - \mu_j|$ in two treatment means, which is of practical significance
3. The probability of a Type II error if any pair of treatments has means that differ by more than $D = |\mu_i - \mu_j|$
4. The variance σ^2

The probability of a Type II error, $\beta(\lambda)$, is determine by using the *noncentral* F distribution with degrees of freedom v_1 and v_2 and the *noncentrality parameter*

$$\lambda = \frac{r\sum_{i=1}^{t}(\mu_i - \mu_.)^2}{\sigma^2}$$

where $\mu_. = \dfrac{1}{t}\displaystyle\sum_{i=1}^{t}\mu_i$. The minimum value of λ for the situation in which at least one pair of treatments has means differing by D units or more is given by

$$\lambda = \frac{rD^2}{2\sigma^2}$$

Table 13 in the Appendix contains the power of the F test, which is the same as $1 - \beta(\lambda)$. The table uses the parameter $\phi = \sqrt{\lambda/t}$ to specify the alternative values of the μ_is. Using this table, we can determine the necessary number of replications to meet the given specifications. The following example will illustrate the requisite calculations.

EXAMPLE 14.14

Refer to Example 14.13, in which a researcher is designing a project to study the yield of pecans under four rates of nitrogen application. The researcher knows that if the average pecan yields differ by more than 15 pounds, there is an economical advantage in using the treatment providing the higher yield. Thus, the researcher wants to determine the number of replications necessary to be 90% certain that the F test will reject H_0 and hence detect a difference in the average yields whenever any pair of nitrogen rates produces average pecan yields differing by more than 15 pounds. The test must have $\alpha = .05$.

Solution From previous experiments, the yields have ranged from 40 pounds to 70 pounds. Thus, an estimate of σ is given by

$$\hat{\sigma} = \frac{70 - 40}{4} = 7.5$$

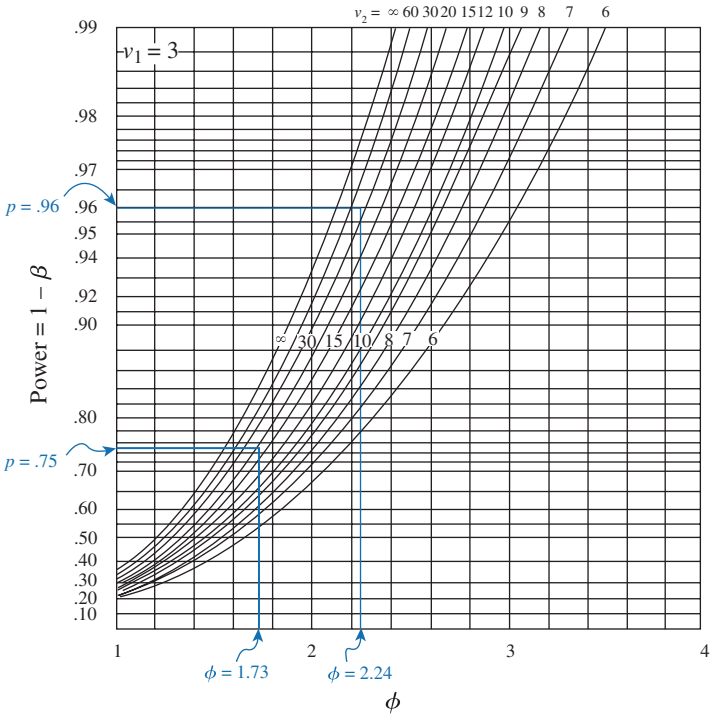


FIGURE 14.13
Power of the analysis of variance test ($\alpha = .05$, $t = 4$)

We have $\alpha = .05, t = 4, v_1 = t - 1 = 4 - 1 = 3$, and $v_2 = N - t = rt - t = t(r - 1) = 4(r - 1)$, where r is the required number of replications. Furthermore, $D = 15$, and, hence,

$$\phi = \sqrt{\frac{rD^2}{2t\hat{\sigma}^2}} = \sqrt{\frac{r(15)^2}{2(4)(7.5)^2}} = .707\sqrt{r}$$

Figure 14.13 contains the power curves needed to solve this problem. Note that $v_1 = 3, \alpha = .05$, and the curves are labeled v_2. We will determine the value of r such that the power is at least .90 when $\phi = .707\sqrt{r}$. We will accomplish this by selecting values of r until we reach the necessary threshold.

The method of determining the proper value for r is by trial and error. First, we guess $r = 6$. Next, we compute $v_2 = 4(6 - 1) = 20$ and $\phi = .707\sqrt{6} = 1.73$. In Figure 14.13, we locate $\phi = 1.73$ on the axis labeled ϕ and draw a vertical line from 1.73 to the curve labeled 20. We then draw a horizontal line to the axis labeled power $= 1 - \beta$ and read the value .75. Thus, if we used six replications in the experiment, our power would only be .75 when $D = 15$, which is too small. We next try $r = 10$ and find that the power is .96. This value would be acceptable; however, a smaller value of r may achieve our goal. Thus, we try $r = 8$ and find that the power equals .89. This value is just slightly too small. Finally, we find that the power is .93 when $r = 9$. Thus, the experiment requires nine replications to meet its specifications. The calculations are summarized in Table 14.30.

TABLE 14.30
Determining the number of replications

r	$v_2 = 4(r - 1)$	$\phi = .707\sqrt{r}$	Power
6	20	1.73	.75
10	36	2.24	.96
8	28	2.00	.89
9	32	2.12	.93

When the experiment has a factorial treatment structure, the calculation of the sample size for an equally replicated design could appear initially to involve rather complex calculations. Suppose we want to test for an interaction between the two factors A and B. This set of hypotheses expressed in terms of the treatment means, μ_{ij}, is

$$H_0: \ \mu_{ij} - \mu_{ik} = \mu_{hj} - \mu_{hk} \text{ for all } (i, j, k, h) \text{ versus}$$

$$H_a: \ \mu_{ij} - \mu_{ik} \neq \mu_{hj} - \mu_{hk} \text{ for at least one set } (i, j, k, h)$$

Reject H_0 if $F = \dfrac{\text{MSAB}}{\text{MSE}} \geq F_{\alpha,\ (a-1)(b-1),\ ab(r-1)}$.

The calculation of the power of this test statistic involves the distribution of a noncentral F distribution with df $= (a - 1)(b - 1), ab(r - 1)$, and noncentrality parameter

$$\lambda = \frac{r}{\sigma_e^2} \sum_{i=1}^{a} \sum_{j=1}^{b} (\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..})^2$$

For specified values of the noncentrality parameter, λ_0, the probability of Type I error, α, and the power of the test, γ_0, determine the minimum value of r such that the power of the test exceeds γ_0 whenever $\lambda \geq \lambda_0$.

Use Table 13 in the Appendix with $\varphi = \sqrt{\lambda/t}$ and $t = ab$ to determine the appropriate value of r. The sample size is then given by $n = rt$.

The above approach is not very realistic because specifying appropriate values for λ_0 is not very intuitive to a researcher, businessperson, or engineer. The following approach follows the methodology used in single-factor experiments.

Determine r by specifying differences in the treatment means:

a. Let $D = \mu_{ij} - \mu_{kh}$ be the difference in any two treatments that the researcher deems important to detect.
b. From our previous results, we know that the minimum value of λ is

$$\lambda = \frac{r \sum_{i=1}^{a} \sum_{j=1}^{b} (\mu_{ij} - \mu_{..})^2}{\sigma_e^2} = \frac{rD_i^2}{2\sigma_e^2}$$

c. Determine the minimum value of r such that the power of the test exceeds γ_0 whenever $\lambda \geq \lambda_0 = rD_i^2/2\sigma_\varepsilon^2$. The result is obtained by using Table 13 in the Appendix, as was done in a single-factor experiment.

After determining the number of replications needed, the number of experimental units may be such that it is physically impossible to conduct the complete experiment at the same time or in the same location. In this type of situation, we can use the concept of randomized complete block designs, with the blocks being either time or location. In Example 14.14, we determined that nine replications of the four treatments or 36 experimental units were needed. Suppose that we had only 12 experimental units at a given location within an agricultural research center. However, there were three such locations, each containing 12 experimental plots. We could thus run three replications of each treatment at each of the three locations. The locations would serve as blocks for the experimental design. We will study the design of randomized block experiments in the next chapter.

EXAMPLE 14.15

An oil pipeline company researcher wishes to study the difference in response times (in milliseconds) for three different types of circuits used in an automatic value shutoff mechanism. There are three major manufacturers of circuits that

will participate in the study. In order to evaluate the difference in the performance of the circuits within each type of circuit from each of the three manufacturers, she decides it is necessary to evaluate r circuits of each type from each of the manufacturers. How large must r be in order to obtain an $\alpha = .05$ test having a power of at least .90 whenever the difference in the mean response times between two of the nine circuits is greater than 2.5 milliseconds? From previous studies, the variation in response times is given by $\sigma_e = 1.1$ milliseconds.

Solution There are $t = (3)(3) = 9$ treatments in this study. The other parameters are given by

$$\hat{\sigma}_e = 1.1 \quad D = 2.5 \quad \gamma_0 = .90 \quad v_1 = t - 1 = 8 \quad v_2 = t(r-1) = 9(r-1)$$

$$\phi = \sqrt{\frac{rD^2}{2t\hat{\sigma}_e^2}} = \sqrt{\frac{r(2.5)^2}{2(9)(1.1)^2}} = .536\sqrt{r}$$

For each value of r, compute v_2 and ϕ, and then obtain the power value from Table 13 ($\alpha = .05, t = 9$) in the Appendix. The results are summarized in Table 14.31.

TABLE 14.31
Determining the number of replications

r	$v_2 = 9(r-1)$	$\phi = .536\sqrt{r}$	Power
3	18	.93	.32
4	27	1.07	.47
5	36	1.20	.62
6	45	1.31	.73
7	54	1.42	.82
8	63	1.52	.88
9	72	1.61	.93

From Table 14.31, the required number of replications is $r = 9$. Thus, the experiment would require $n = tr = abr = (3)(3)(9) = 81$ experimental units to achieve the specified requirements. ■

14.7 RESEARCH STUDY: Development of a Low-Fat Processed Meat

In Section 14.1, we described a research study in which meat scientists investigated methods by which a variety of low-fat meat products could be developed that maintained product yields and minimized formulation costs while retaining acceptable palatability. The researchers determined that lowering the cost of production without affecting the quality of the low-fat meat product required the substitution of nonmeat ingredients such as soy protein isolates (SPI) for a portion of the meat block. When replacing meat with SPI, it is necessary to incorporate konjac flour into the product to maintain the appealing characteristics of high-fat products.

Designing the Data Collection

The three factors identified for study were the type of konjac blend, amount of konjac blend, and percentage of SPI substitution in the meat product. There were many other possible factors of interest, such as cooking time, temperature, type of meat product, and length of curing. However, the researchers selected the commonly used levels of these factors in a commercial preparation of bologna and narrowed the study to the three most important factors. This resulted in an experiment having 12 treatments, as displayed in Table 14.32.

TABLE 14.32
Mean values for meat texture in low-fat bologna study

Konjac Level (%)	Konjac Blend	SPI (%)	Texture Readings	Mean Texture
.5	KSS	1.1	107.3, 110.1, 112.6	110.0
.5	KSS	2.2	97.9, 100.1, 102.0	100.0
.5	KSS	4.4	86.8, 88.1, 89.1	88.0
.5	KNC	1.1	108.1, 110.1, 111.8	110.0
.5	KNC	2.2	108.6, 110.2, 111.2	110.0
.5	KNC	4.4	95.0, 95.4, 95.5	95.3
1	KSS	1.1	97.3, 99.1, 100.6	99.0
1	KSS	2.2	92.8, 94.6, 96.7	94.7
1	KSS	4.4	86.8, 88.1, 89.1	88.0
1	KNC	1.1	94.1, 96.1, 97.8	96.0
1	KNC	2.2	95.7, 97.6, 99.8	97.7
1	KNC	4.4	90.2, 92.1, 93.7	92.0

The objective of this study was to evaluate various types of konjac blends as a partial lean meat replacement and to characterize its effect in a very low-fat bologna model system. Two types of konjac blends (KSS = konjac flour/starch and KNC = konjac flour/carrageenan/starch), at levels .5% and 1%, and three meat protein replacement levels with SPI (1.1, 2.2, and 4.4%) were selected for evaluation.

The experiment was conducted as a completely randomized design with a $2 \times 2 \times 3$ three-factor factorial treatment structure and three replications of the 12 treatments. There were a number of response variables measured on the 36 runs of the experiment, but we will discuss the results for the texture of the final product as measured by an Instron universal testing machine. The responses and their means are given in Table 14.32.

Analyzing the Data

Because the number of calculations needed to obtain the sum of squares in a three-factor experiment is substantial and consequently may lead to significant round-off error, we will use a software program to obtain the results shown in Table 14.33.

TABLE 14.33
AOV table for data in case study, a three-factor factorial experiment

```
General Linear Models Procedure

Dependent Variable: Texture of Meat:

                                  Sum of          Mean
Source                    DF     Squares        Square    F Value     Pr > F
Model                     11  2080.28750     189.11705      62.40     0.0001
Error                     24    72.74000       3.03083
Corrected Total           35  2153.02750

                  R-Square         C.V.      Root MSE              Y Mean
                  0.966215     1.769387       1.74093             98.3917

Source                    DF   Type III SS   Mean Square   F Value     Pr > F
Main Effects:
L                          1    526.70250     526.70250    173.78     0.0001
B                          1    113.42250     113.42250     37.42     0.0001
P                          2   1090.11500     545.05750    179.84     0.0001
Interactions:
L*B                        1     44.22250      44.22250     14.59     0.0008
L*P                        2    182.53500      91.26750     30.11     0.0001
B*P                        2    115.84500      57.92250     19.11     0.0001
L*B*P                      2      7.44500       3.72250      1.23     0.3106
```

TABLE 14.34
Table of means for data in case study

Level (%)	Blend	SPI (%)	Two-Way Means
.5	KSS	*	99.3
.5	KNC	*	105.1
1	KSS	*	93.9
1	KNC	*	95.2
.5	*	1.1	110.0
.5	*	2.2	105.0
.5	*	4.4	91.7
1	*	1.1	97.5
1	*	2.2	96.2
1	*	4.4	90.0
*	KSS	1.1	104.5
*	KSS	2.2	97.4
*	KSS	4.4	88.0
*	KNC	1.1	103.0
*	KNC	2.2	103.9
*	KNC	4.4	93.7

The notation in the AOV table is as follows: L refers to the konjac level, B refers to the type of konjac blend, and P refers to the level of SPI. Since three-way interaction in the AOV model was not significant (L*B*P, $p = .3106$), we next examine the two-way interactions. The three sets of two-way interactions had the following levels of significance: L*B, $p = .0008$ L*P, $p < .0001$ and B*P, $p < .0001$. Thus, all three were highly significant. To examine the types of relationships that may exist among the three factors, we need to obtain the sample means, $\bar{y}_{ij.}$, $\bar{y}_{i.k.}$, and $\bar{y}_{.jk.}$. These values are given in Table 14.34.

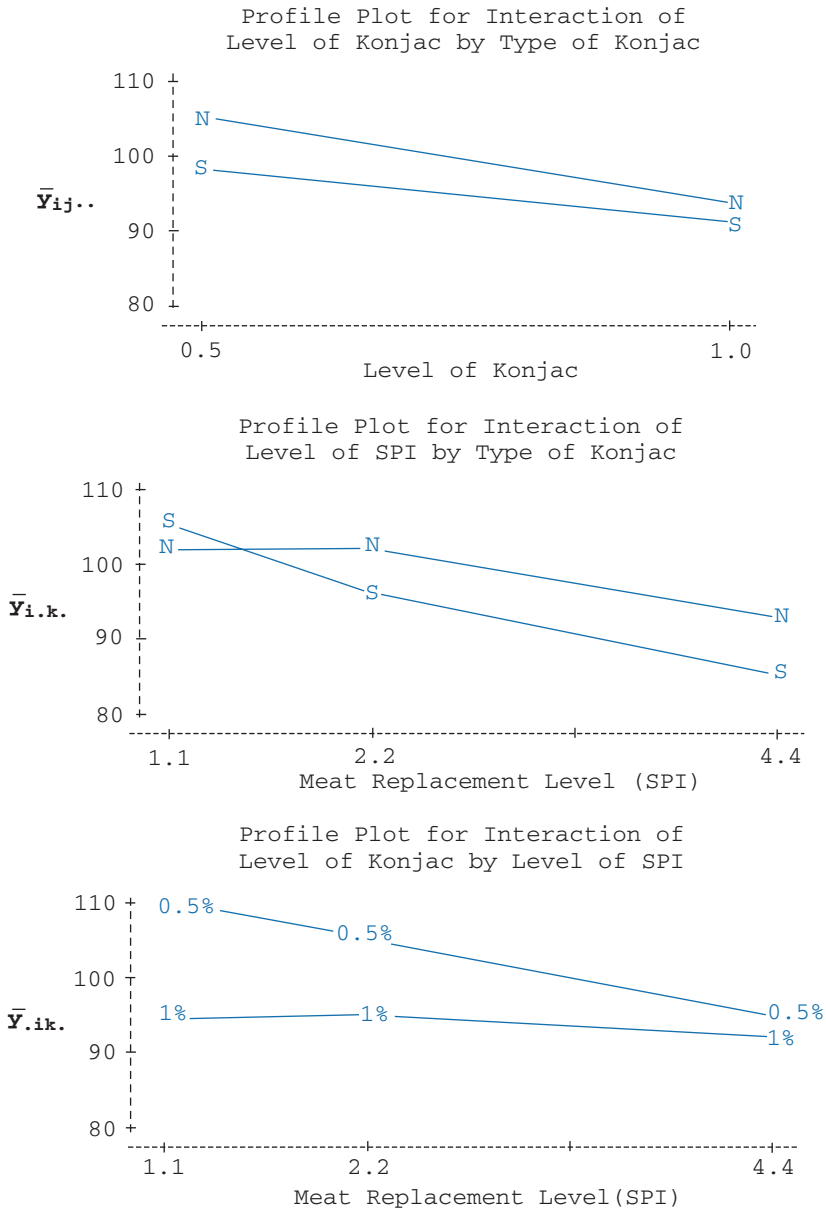
The means in the table are then plotted in Figure 14.13 to yield the profile plots for the two-way interactions of level of konjac with type of konjac, level of konjac with level of SPI, and type of konjac with level of SPI.

From Figure 14.14, we can observe that there are considerable differences in the mean texture of the meat product depending on the type of konjac, the level of konjac, and the level of SPI in the meat product. When the level of konjac is 1%, there is very little difference in the mean textures of the meat; however, at the .5% level, the KNC blend of konjac produced a product with a higher mean texture than did the KSS blend of konjac. When considering the effect of level of SPI on the mean texture of the bologna, we can observe that at a level of 1.1% SPI, there was a sizable difference between using .5% konjac and 1% konjac. As the level of SPI increased, the size of the difference decreased markedly. Furthermore, at a 1.1% level of SPI, there was essentially no difference between the two blends of konjac, but as the level of SPI increased, the KNC blend produced a meat product having a higher texture than the KSS blend. These observations about the relationships among the three factors and the mean textures of the meat product need to be confirmed using multiple-comparison procedures, which will be done after an analysis of the residuals.

Figure 14.15 contains the residuals analysis for the texture data. We obtain the residuals using the formula

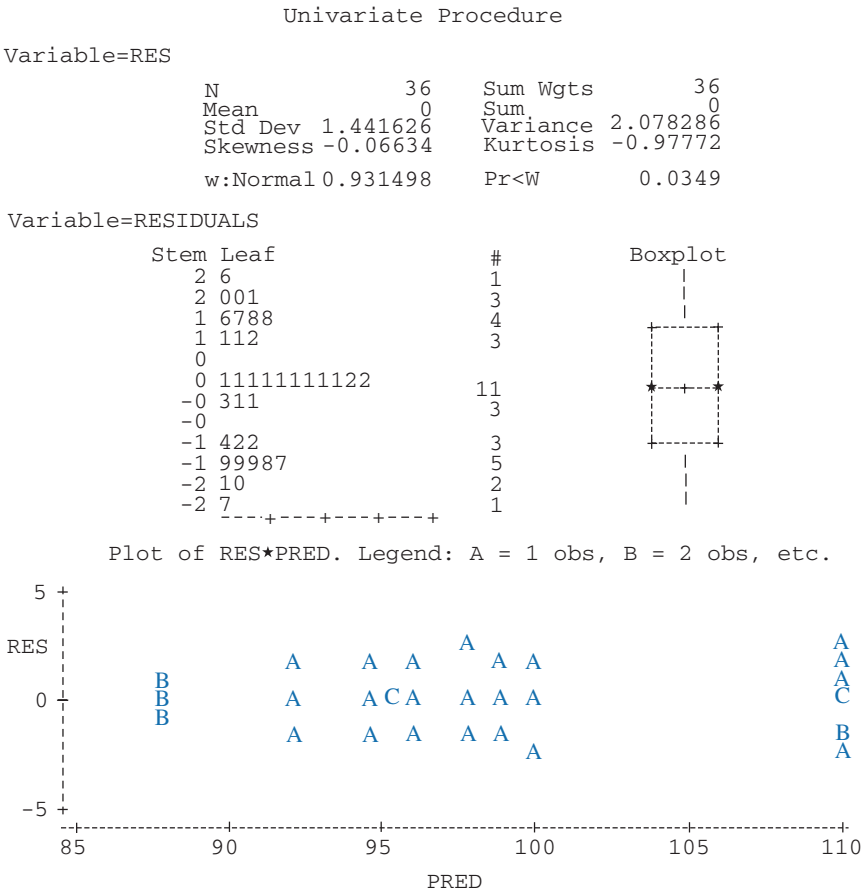
$$e_{ijkm} = y_{ijkm} - \bar{y}_{ijk.}$$

FIGURE 14.14
Profile plots of the two-way interactions



An examination of the stem-and-leaf plot and boxplot reveals that the residuals are nearly symmetric but have a sharp peak near 0. The Shapiro–Wilk test for normality has a p-value of .0349, which reflects the somewhat nonnormal nature of the residuals. However, because there are no outliers and very few residuals even near extreme in size, the normality assumption is nearly met. The plot of the residuals versus the estimated treatment means $\bar{y}_{ijk.}$ reveals a slight increase in variability as the mean texture readings increased. However, this increase is not large enough to overcome the natural robustness of the F test for small deviations from the model conditions. Thus, both the normality and the equal variance conditions appear to be satisfied, and we would conclude that the F tests in the Table 14.33 would be valid.

FIGURE 14.15
Residuals analysis for case study



Because the three-way interaction, L*B*P, was not significant (p-value = .3106), we will examine the two-way interactions of interest to the researchers. They wanted to investigate the effect on mean texture of increasing the percentage of SPI in the meat product. Thus, we need to examine the differences in mean texture as a function of the percentage of SPI. Because there was a significant (p-value $< .0001$) interaction between SPI and level of konjac and a significant (p-value $< .0001$) interaction between SPI and type of konjac, we need to conduct four different mean separations of the levels of the percentage of SPI.

First, we will compare the mean textures across the percentage of SPI separately for each of the two values of level of konjac: 0.5% and 1.0%. The value of Tukey's W is given by

$$W = q_\alpha(t, \text{df}_{\text{error}}) \sqrt{\frac{s_\varepsilon^2}{n_t}}$$

where $t = 3$, the number of levels of the percentage of SPI, $\text{df}_{\text{error}} = 24$, $s_\varepsilon^2 = 3.0308$ from Table 14.33, and $n_t = 6$, the number of observations in each of the percentage of SPI means at each of the values of level of konjac because $\bar{y}_{i.k.}$ is based on six data values. Thus, from Table 10 in the Appendix, we find $q_\alpha(t, \text{df}_{\text{error}}) = q_{.05}(3, 24) = 3.53$, which yields

$$W = 3.53\sqrt{\frac{3.0308}{6}} = 2.51$$

TABLE 14.35
Mean texture across levels of the percentage of SPI at each level of konjac

	SPI (%)		
Level of Konjac	**1.1**	**2.2**	**4.4**
0.5%	110.0	105.0	9.17
	a	b	c
1.0%	97.5	96.2	90.0
	a	a	b

TABLE 14.36
Mean texture across levels of the percentage of SPI for each konjac blend

	SPI (%)		
Konjac Blend	**1.1**	**2.2**	**4.4**
KSS	104.5	97.4	88.0
	a	b	c
KNC	103.0	103.9	93.7
	a	a	b

Thus, any pair of means $\bar{y}_{i.k.}$ and $\bar{y}_{i.k'.}$ that differ by more than 2.51 will be declared to be significantly different at the $\alpha = .05$ level. A summary of results is given in Table 14.35.

For the 0.5% level of konjac, all three percentages of SPI yield significantly different mean textures; the higher the level of the percentage of SPI, the lower the value for mean texture. For the 1.0% level of konjac, the 1.1 and 2.2 percentages of SPI have nonsignificantly different mean textures, whereas the 4.4 percentage of SPI has a significantly lower mean texture in comparison to the 1.1 and 2.2 percentages. Thus, the relationship between the percentage of SPI and mean texture is different at the two levels of konjac. Similarly, we obtain the following results (Table 14.36) for the relationship between mean texture and the percentage of SPI at the two blends of konjac. The values of all the quantities in W remain the same as before, because the number of observations in each of the type of konjac– percentage of SPI means, $\bar{y}_{ik}$, is $n_t = 6$. Thus, $W = 2.51$.

For the KSS blend, all three percentages of SPI yield significantly different mean textures. For the KNC blend, the 1.1 and 2.2 percentages of SPI have nonsignificantly different mean textures, whereas the 4.4 percentage of SPI has a significantly lower mean texture in comparison to the 1.1 and 2.2 percentages. Thus, the relationship between percentage of SPI and mean texture is different for the blends of konjac.

14.8 Summary and Key Formulas

In this chapter, we discussed the analysis of variance for various treatment structures in a completely randomized design. Included were single-factor, two-factor, and three-factor treatment structures. The factorial treatment structure is useful in investigating the effect of one or more factors on an experimental response. The crucial motivation in using factorial treatment structures is to determine whether or not an interaction exists between the factors.

For each of the treatment structures discussed in this chapter, we presented a description of the design layout (including the arrangement of treatments), a model, and the analysis of variance. We also discussed how one could conduct multiple comparisons between treatment means for both a single factor and a

multiple-factor treatment structure. Finally, a method for determining the appropriate number of replications to achieve specified design criteria was presented.

balanced design

For the most part, the development of the analysis of variance and the determination of replication size were for a **balanced design**—that is, a design in which each treatment (factor–level combination) is randomly assigned to the same number of experimental units. It is only in a balanced design that explicit formulas for the various sums of squares can be displayed. When the design is unbalanced, the methodology for obtaining the sums of squares is more complex and, in most cases, should be computed using an appropriate statistical software program.

Key Formulas

1. One factor in a completely randomized design

   Model: $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

   Sum of Squares (Equal replications):

   Total $\quad \text{TSS} = \Sigma_{ij}(y_{ij} - \bar{y}_{..})^2$

   Treatment $\quad \text{SST} = n\,\Sigma_i(\bar{y}_{i.} - \bar{y}_{..})^2$

   Error $\quad \text{SSE} = \Sigma_{ij}(e_{ij})^2 = \Sigma_{ij}(y_{ij} - \bar{y}_{i.})^2 = \text{TSS} - \text{SST}$

2. Two-factor factorial treatment structure in a completely randomized design

   Model: $y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$

   Sum of Squares (Equal replications):

   Total $\quad \text{TSS} = \Sigma_{ijk}(y_{ijk} - \bar{y}_{...})^2$

   Factor A $\quad \text{SSA} = bn\,\Sigma_i(\bar{y}_{i..} - \bar{y}_{...})^2$

   Factor B $\quad \text{SSB} = an\,\Sigma_j(\bar{y}_{.j.} - \bar{y}_{...})^2$

   Interaction $\quad \text{SSAB} = n\,\Sigma_{ij}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$

   Error $\quad \text{SSE} = \Sigma_{ijk}(y_{ijk} - \bar{y}_{ij.})^2 = \text{TSS} - \text{SSA} - \text{SSB} - \text{SSAB}$

3. $100(1 - \alpha)\%$ simultaneous confidence interval for difference in t treatment means

$$(\bar{y}_i - \bar{y}_{i'}) \pm \frac{q_\alpha(t, \nu)}{\sqrt{2}}\sqrt{s_\varepsilon^2\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}$$

14.9 Exercises

14.2 Completely Randomized Design with a Single Factor

Edu.

14.1 Researchers in child development are interested in developing ways to increase the spatial–temporal reasoning of preschool children. Spatial–temporal reasoning relates to the child's ability to visualize spatial patterns and mentally manipulate them over a time-ordered sequence of spatial transformations. This ability, often referred to as thinking in pictures, is important for generating and conceptualizing solutions to multistep problems and is crucial in early child development. The researchers want to design a study to evaluate which of several methods proposed to accelerate the growth in spatial–temporal reasoning yields the greatest increase in a child's development in this area. There are three methods proposed: taking piano lessons for 3 months, playing specially developed computer video games for 3 months, and playing specially designed games in small groups supervised by a trained instructor. The researchers measure the effectiveness of the three programs by assessing the children and assigning each one a reasoning score both before and after participation in the program. The difference in these two scores is the

response variable. A control group is also included to measure the change in reasoning for children not given any special instruction. A pilot study with only 20 students was to be conducted prior to the complete study to determine potential problems. Demonstrate how to assign 5 of the 20 students to each of the four instructionalonditions—no instruction (control), piano lessons, computer video games, and instructor—so that the assignment is completely random.

14.2 Refer to Exercise 14.1. The researchers decide to use the following model, which relates the response variable y to the four instructional conditions.

$$y = \mu + \tau_i + \varepsilon_{ij} \quad \text{for} \quad i = 1, 2, 3, 4 \quad \text{and} \quad j = 1, 2, 3, 4, 5$$

a. Write an equation relating the mean reasoning score, μ_i, to the parameters in the above model without any constraints on the model parameters.

b. Rewrite the equation relating the mean reasoning score, μ_i, to the parameters in the above model after imposing the standard constraints placed on the model parameters.

14.3 Refer to Exercise 14.1. After running the pilot study, the researchers conduct a study involving 100 students. Twenty-five students were randomly assigned to each of the four instructional conditions. The data are given here.

a. Conduct an analysis of variance, and summarize your results in an AOV table.

b. Test the research hypothesis that there is a difference in the effectiveness means of the methods of instruction. Use $\alpha = .05$.

c. Apply a multiple-comparison procedure to determine pairwise differences in the three instructional methods. Use $\alpha = .05$.

d. Was there significant evidence that all three methods of instruction produced higher mean reasoning scores than the mean reasoning score for the control?

		Method of Instruction		
Student	**Control**	**Piano**	**Computer**	**Instructor**
1	−3.4	−.2	7.7	12.0
2	−2.8	5.2	5.5	4.1
3	2.2	6.6	−.8	5.9
4	−.8	5.2	7.4	13.5
5	2.8	−.6	.1	7.5
6	−5.9	5.4	11.7	9.3
7	7.8	3.1	1.2	7.1
8	−3.5	6.5	3.8	−.9
9	2.9	2.4	5.1	8.3
10	1.9	6.2	4.3	9.8
11	−.2	7.9	3.9	11.1
12	1.5	7.9	6.9	4.9
13	.4	6.6	2.8	5.8
14	−.5	.2	5.4	2.8
15	1.1	1.9	2.5	12.0
16	5.3	1.3	5.2	8.6
17	−4.0	1.8	3.1	2.0
18	−1.3	3.1	6.6	5.9
19	2.6	1.4	.2	5.6
20	−.9	2.1	7.1	11.6
21	−.6	6.6	9.2	7.8
22	−5.0	7.0	3.0	7.2
23	2.4	−.7	2.3	8.3
24	−.1	4.1	10.2	6.5
25	−4.7	3.8	4.7	8.3

14.4 In order for the conclusions reached in Exercise 14.3 to be valid, the conditions of normality, equal variance, and independence must be satisfied. Use the residuals from the fitted model to assess the three conditions. (Refer to the discussion in Section 8.4.)

- **a.** Was there significant evidence of a violation of the normality condition?
- **b.** Was there significant evidence that the variance in reasoning scores was different for the three methods and the control?
- **c.** What is the justification for concluding that the 100 reasoning scores are independent?
- **d.** If the condition of normality and/or equal variance is violated, what are some alternative methods of analysis?

Engin. **14.5** The production manager of a large casting firm is studying different methods to increase productivity in the workforce of the company. The process engineer and personnel in the human resources department develop three new incentive plans (plans B, C, and D) and design a study to compare these incentive plans with the current plan (plan A). Twenty workers are randomly assigned to each of the four plans. The response variable is the total number of units produced by each worker during 1 month on the incentive plan. The data are given in the following table.

	Incentive Plan			
Worker	**A**	**B**	**C**	**D**
1	422	521	437	582
2	431	545	422	639
3	784	600	473	735
4	711	406	478	800
5	641	563	397	853
6	709	361	944	748
7	344	387	394	622
8	599	700	890	514
9	511	348	488	714
10	381	944	521	627
11	349	545	387	548
12	387	337	633	644
13	394	427	627	736
14	621	771	444	528
15	328	752	1,467	595
16	636	810	828	572
17	388	406	644	627
18	901	537	1,154	546
19	394	816	430	701
20	350	369	508	664
Mean	514.1	557.2	628.3	649.8
St Dev	171.8	184.4	290.2	93.1

- **a.** State the null and alternative hypotheses being tested by the F statistic in the AOV table.
- **b.** Is there significant evidence ($\alpha = .05$) that the mean output associated with the four incentive plans is different?
- **c.** Use Tukey's W procedure to identify the pairs of incentive plans that have different output means.

14.6 In order for the conclusions reached in Exercise 14.5 to be valid, the conditions of normality, equal variance, and independence must be satisfied. Use the residuals from the fitted model to assess the three conditions. Refer to the discussion in Section 8.4.

- **a.** Is there significant evidence of a violation of the normality condition?
- **b.** Is there significant evidence that the variances in reasoning scores were different for the three methods and the control?
- **c.** What is the justification for concluding that the 100 reasoning scores are independent?
- **d.** If the condition of normality and/or equal variance is violated, what are some alternative methods of analysis?

14.7 Refer to Exercise 14.5. When the normality condition is violated, an alternative to the F test is the Kruskal–Wallis test (see Section 8.6).

- **a.** Test for differences in the median outputs of the four incentive plans. Use $\alpha = .05$.
- **b.** Why do you think the conclusions reached using the Kruskal–Wallis test differ from the conclusions reached using the F test from the AOV table in Exercise 14.5?

14.3 Factorial Treatment Structure

Bus.

14.8 A large advertising firm specializes in creating television commercials for children's products. The firm wants to design a study to investigate factors that may affect the lengths of time a commercial is able to hold a child's attention. A preliminary study determines that two factors that may be important are the age of the child and the type of product being advertised. The firm wants to determine whether there were large differences in the mean length of time that the commercial is able to hold the child's attention depending on these two factors. If there proves to be a difference, the firm would then attempt to determine new types of commercials depending on the product and targeted age group. Three age groups are used:

A_1: 5–6 years A_2: 7–8 years A_3: 9–10 years

The types of products selected are

P_1: breakfast cereals P_2: video games

A group of 30 children is recruited in each age group, and 10 are randomly assigned to watch a 60-second commercial for each of the two products. Researchers record their attention spans during the viewing of the commercial. The data are given here.

Child	A_1–P_1	A_2–P_1	A_3–P_1	A_1–P_2	A_2–P_2	A_3–P_2
1	19	19	37	39	30	51
2	36	35	6	18	47	52
3	40	22	28	32	6	43
4	30	28	4	22	27	48
5	4	1	32	16	44	39
6	10	27	16	2	26	33
7	30	27	8	36	33	56
8	5	16	41	43	48	43
9	34	3	29	7	23	40
10	21	18	18	16	21	51
Mean	22.9	19.6	21.9	23.1	30.5	45.6

Mean by age group:	A_1	A_2	A_3	Mean by product type:	P_1	P_2
	23.0	25.05	33.75		21.47	33.07

- **a.** Identify the design.
- **b.** Write a model for this situation, identifying all the terms in the model.
- **c.** Estimate the parameters in the model.
- **d.** Compute the sum of squares for the data, and summarize the information in an AOV table.

14.9 Refer to Exercise 14.8.

a. Draw a profile plot for the two factors, age and product type.

b. Perform appropriate F tests and draw conclusions from these tests concerning the effects of age and product type on the mean attention spans of the children.

14.10 Refer to Exercise 14.8.

Use residual plots to determine whether any of the conditions required for the validity of the F tests have been violated.

Bus. 14.11 Commercially produced ice cream is made from a mixture of ingredients:

- A minimum of 10% milk fat
- 9–12% milk solids: this component, also known as the serum solids, contains the proteins (caseins and whey proteins) and carbohydrates (lactose) found in milk
- 12–16% sweeteners: usually a combination of sucrose and/or glucose-based corn syrup sweeteners
- 0.2–0.5% stabilizers and emulsifiers—e.g., agar or carrageenan extracted from seaweed
- 55%–64% water, which comes from milk solids or other ingredients

Air is incorporated with the above ingredients during the mixing process. Less-expensive ice creams contain lower-quality ingredients, and more air is incorporated during the mixing process. The finest ice creams have between 3% and 15% air. Because most ice cream is sold by volume, it is economically advantageous for producers to reduce the density of the product in order to cut costs. A food scientist is investigating how varying the amounts of the above ingredients impacts the sensory rating of the final product. The scientist decides to use three levels of milk fat: 10%, 12%, 15%; three amounts of air: 5%, 10%, 15%; and two levels of sweeteners: 12%, 16%. Three replications of each of the formulations were produced and the sensory ratings (0–40) obtained; a higher number implies a more favorable sensory rating. The data are given here.

	Sweetener					
	12%			16%		
	Milk Fat			Milk Fat		
Air	10%	12%	15%	10%	12%	15%
	23	27	31	24	38	34
5%	24	28	32	23	36	36
	25	26	29	28	35	39
	36	34	33	37	34	34
10%	35	38	34	39	38	36
	36	39	35	35	36	31
	28	35	26	26	36	28
15%	24	35	27	29	37	26
	27	34	25	25	34	24

a. Identify the design and treatment structure for this study.

b. Write a model for this study, identifying all the terms in the model.

c. For each of the two levels of sweetener, draw profile plots of the effects of the percentages of air and milk fat on the sensory rating of ice cream.

d. From the profile plots, does there appear to be a three-way interaction among the effects of the percentages of sweetener, air, and milk fat on the mean sensory ratings?

14.12 Refer to the study described in Exercise 14.11.

- **a.** Perform appropriate F tests and draw conclusions from these tests concerning the effects of the percentages of sweetener, air, and milk fat on the sensory ratings. Use $\alpha = .05$.
- **b.** Are the conclusions from the F tests consistent with your observations from the profile plots?

14.13 Refer to the study described in Exercise 14.11. Use the residuals from the fitted model to answer the following questions.

- **a.** Is there significant evidence that the residuals have a nonnormal distribution?
- **b.** Is there significant evidence that the residuals do not have constant variances?
- **c.** How could we assess whether or not the residuals are independently distributed?

14.5 Estimation of Treatment Differences and Comparisons of Treatment Means

14.14 Refer to the study described in Exercise 14.8. Use Tukey's W procedure to identify significant differences in the means.

- **a.** Use Tukey's W procedure to identify significant differences in the mean attention spans of the three age groups of children.
- **b.** Use Tukey's W procedure to identify significant differences in the mean attention spans for the types of products.
- **c.** Are your conclusions in part (a) the same for both types of products?

14.15 Refer to the study described in Exercise 14.11.

- **a.** Use Tukey's W procedure to identify significant differences in the mean sensory ratings of the three levels of percentage of milk fat.
- **b.** Use Tukey's W procedure to identify significant differences in the mean sensory ratings of the three levels of percentage of air.
- **c.** Which combination of percentage of milk fat, air, and sweetener appears to yield the highest mean sensory rating?

14.6 Determining the Number of Replications

Edu. **14.16** A state legislature mandates the each school district in its state must conduct an audit of the performance of the district's students on the state reading exam. The purpose is to determine if there are any extreme increases in the individual schools in the district. There are currently four software programs that are capable of conducting the audits with varying degrees of efficiency. The state board of education hires an analyst to design a study to evaluate each of the software programs. The study will involve a random sample of schools running the software on their records. One of the metrics in the evaluation will be the amount of time that the software takes to complete the audit. From the application of the software in other states, the standard deviation in the time to complete the audit was 122.5 minutes. Determine how many schools are required in the study for each software program in order to be able to detect a difference in any pair of software programs of 5 hours using a level .05 test with a power of 90%.

Edu. **14.17** A researcher seeks funding for a study from a federal agency. The study will involve the evaluation of three factors, each having two levels. From the literature, the researcher approximates the standard deviation in the responses to be approximately 9 units. How many experimental units should be included in the budget for the study so that a difference of 20 units or more in any pair of treatment means will be detected with a probability of .80 using an $\alpha = .05$ test?

14.18 Refer to the study described in Exercise 14.8. Determine the number of replications needed to obtain an $\alpha = .05$ test having power of at least 80% that detects a difference of 10 in any pair of treatment means. Use the data from Exercise 14.8 to estimate the value of σ_e^2.

14.19 Refer to the study described in Exercise 14.11. Suppose a new study is to be designed in which only three levels of milk fat and three levels of air will be used. Determine the number of replications needed to obtain an $\alpha = .05$ test having a power of at least 90% that detects a difference of 5 in any pair of treatment means. Use the data from Exercise 14.11 to estimate the value of σ_e^2.

Supplementary Exercises

Ag. **14.20** A study was conducted to compare the effect of four manganese rates (from $MnSO_4$) and four copper rates (from $CuSO_4\ 5H_2O$) on the yield of soybeans. A large field was subdivided into 32 separate plots. Two plots were randomly assigned to each of the 16 factor–level combinations (treatments) and the treatments were applied to the designated plots. Soybeans were then planted over the entire field in rows 3 feet apart. The yields from the 32 plots are given here (in kilograms/hectare).

	Mn				
Cu	**20**	**50**	**80**	**110**	**Cu Mean**
	1,558	2,003	2,490	2,830	
1	1,578	2,033	2,470	2,810	2,221.5
	1,590	2,020	2,620	2,860	
3	1,610	2,051	2,632	2,841	2,278.0
	1,558	2,003	2,490	2,830	
5	1,550	2,010	2,690	2,910	2,255.1
	1,328	2,010	2,887	2,960	
7	1,427	2,031	2,832	2,941	2,302.0
Mn Mean	1,524.9	2,020.1	2,638.9	2,872.8	2,264.2

a. Identify the design for this experiment.
b. Write an appropriate statistical model for this experiment.
c. Construct a profile plot and describe what this plot says about the effect of Mn and Cu on soybean yield.

14.21 Refer to Exercise 14.20.
a. Test for an interaction between the effects of Mn and Cu on soybean yield. Use $\alpha = .05$.
b. What level of Mn appears to produce the highest yield?
c. What level of Cu appears to produce the highest yield?
d. What combination of Cu–Mn appears to produce the highest yield?

14.22 Suppose we have a completely randomized three-factor factorial experiment with levels $3 \times 4 \times 6$, with three replications of each of the 72 treatments. Assume that the three-way interaction is not significant.
a. Write a model to describe the response y_{ijkm} for this type of experiment.
b. Provide a complete AOV table for this type of experiment.
c. Sketch three profile plots to depict the following three two-way interactions: $F_1{*}F_2$ significant but orderly, $F_2{*}F_3$ nonsignificant, and $F_1{*}F_3$ significant and disorderly.

Ag. **14.23** An experiment was set up to compare the effects of different soil pH and calcium additives on the increase in trunk diameters for orange trees. Elemental sulfur, gypsum, soda ash, and other ingredients were applied annually to provide pH value levels of 4, 5, 6, and 7. Three levels of a calcium supplement (100, 200, and 300 pounds per acre) were also applied. All factor–level

combinations of these two variables were used in the experiment. At the end of a 2-year period, three diameters were examined at each factor–level combination. The data appear next.

pH Value	Calcium 100	200	300
4.0	5.2	7.4	6.3
	5.9	7.0	6.7
	6.3	7.6	6.1
5.0	7.1	7.4	7.3
	7.4	7.3	7.5
	7.5	7.1	7.2
6.0	7.6	7.6	7.2
	7.2	7.5	7.3
	7.4	7.8	7.0
7.0	7.2	7.4	6.8
	7.5	7.0	6.6
	7.2	6.9	6.4

a. Construct a profile plot. What do the data suggest?
b. Write an appropriate statistical model.
c. Perform an analysis of variance and identify the experimental design. Use $\alpha = .05$.

14.24 Refer to Exercise 14.23.
a. Test for interactions and main effects. Use $\alpha = .05$.
b. What can you conclude about the effects of pH and calcium on increases in mean trunk diameter for orange trees?

14.25 Refer to Exercise 14.23.
a. Use Tukey's W procedure to determine differences in mean increases in trunk diameter among the three calcium rates. Use $\alpha = .05$.
b. Are your conclusions about the differences in mean increases in diameter among the three calcium rates the same for all four pH values?

14.26 Refer to Exercise 14.23.
a. Use residual analysis to determine whether any of the conditions required to conduct an appropriate F test have been violated.
b. If any of the conditions have been violated, suggest ways to overcome these difficulties.

Med.

14.27 Researchers conducted an experiment to compare the average oral body temperatures for persons taking one of nine different medications often prescribed for high blood pressure. The researchers were concerned that the effect of the drug may be different depending on the severity of the patient's high blood pressure disorder. Patients with high blood pressure who satisfied the study's entrance criteria were classified into one of the three levels of severity of the blood pressure disorder. The patients were then randomly assigned to receive one of the nine medications. Each patient in the study was given the assigned medication at 6:00 A.M. of the designated study day. Temperatures were taken at hourly intervals beginning at 8:00 A.M. and continuing for 10 hours. During this time, the patients were not allowed to do any physical activity and had to lie in bed. To eliminate the variability of temperature readings within a day, the average of the hourly determinations was the recorded response for each patient. These data are given in the accompanying table.
a. Identify the design for this experiment.
b. Write an appropriate statistical model and identify the parameters of the model.

	Medication								
Severity	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
	97.8	98.1	98.0	97.3	97.9	97.9	97.1	98.0	97.8
	97.2	98.1	97.8	97.3	97.8	97.9	97.6	97.8	98.0
1	97.6	98.0	98.1	97.5	97.8	97.8	97.3	98.0	97.7
	97.2	97.7	97.8	97.5	97.7	97.8	97.7	97.9	97.9
	97.6	97.7	97.9	97.6	97.8	97.6	97.5	98.0	97.8
	97.6	97.8	97.9	97.5	97.8	98.0	97.6	97.9	98.0
	97.4	97.7	98.1	97.4	97.8	97.7	97.5	98.0	97.6
2	97.3	97.6	97.8	97.5	97.7	97.8	97.6	97.9	98.0
	97.5	97.7	97.8	97.6	97.7	97.9	97.5	97.9	97.9
	97.5	97.7	97.6	97.7	97.8	97.8	97.3	97.8	97.9
	97.5	97.6	98.0	97.9	97.7	97.9	97.4	97.8	98.0
	97.9	97.7	97.8	97.8	97.8	98.0	97.8	97.8	98.1
3	97.6	97.9	98.1	97.8	97.9	97.7	97.4	98.0	97.9
	97.6	97.9	97.7	97.8	98.0	97.9	97.6	97.9	98.1
	97.7	97.8	98.7	97.6	98.1	97.9	97.6	97.8	97.9

14.28 Refer to Exercise 14.27.

a. Construct an AOV table for the experiment.

b. Are the differences in mean temperatures for the nine medications the same for all three severities of the blood pressure disorders? Use $\alpha = .05$.

c. Is there a significant difference in mean temperatures for medications and severity of the disorder? Use $\alpha = .05$.

d. Use a profile plot to assist in discussing your conclusions concerning the effects of medication and severity on the mean temperatures of the patients.

Med.

14.29 A physician was interested in examining the relationship between the work performed by individuals in an exercise tolerance test and the excess weight (as determined by standard weight–height tables) they carried. To do this, a random sample of 28 healthy adult females, ranging in age from 25 to 40, was selected from the community clinic during routine visits for physical examinations. The selection process was restricted so that seven persons were selected from each of the following weight classifications:

Normal weight (less than 10% underweight)
1%–10% overweight
11%–20% overweight
More than 20% overweight

As part of the physical examination, each person was required to exercise on a bicycle ergometer until the onset of fatigue. The time to fatigue (in minutes) was recorded for each person. The data are given next.

Classification	**Fatigue Time**
Normal	25, 28, 19, 27, 23, 30, 35
1%–10% overweight	24, 26, 18, 16, 14, 12, 17
11%–20% overweight	15, 18, 17, 25, 12, 10, 23
More than 20% overweight	10, 9, 18, 14, 6, 4, 15

a. Identify the experimental design and write an appropriate statistical model.

b. Use $\alpha = .05$ and perform an analysis of variance.

14.30 Refer to Exercise 14.29.

a. How would you design an experiment to investigate the effects of age, gender, and excess weight on fatigue time?

b. Suppose the physician wanted to investigate the relationship among the quantitative variables percentage overweight, age, and fatigue time. Write a possible model.

Env. **14.31** An experiment was conducted to investigate the heat loss for five different designs for commercial thermal panes. The researcher, in order to obtain results that would be applicable throughout most regions of the country, decided to evaluate the panes at five temperatures: 0°F, 20°F, 40°F, 60°F, and 80°F. A sample of 10 panes of each design was obtained. Two panes of each design were randomly assigned to each of the five exterior temperature settings. The interior temperature of the test was controlled at 70°F for all five exterior temperatures. The heat losses associated with the five pane designs are given here.

	Pane Design				
Exterior Temperature Setting (°F)	**A**	**B**	**C**	**D**	**E**
80	7.2, 7.8	7.1, 7.9	8.1, 8.8	8.3, 8.9	9.3, 9.8
60	8.1, 8.1	8.0, 8.9	8.2, 8.9	8.1, 8.8	9.2, 9.9
40	9.0, 9.9	9.2, 9.8	10.0, 10.8	10.2, 10.7	9.9, 9.0
20	9.2, 9.8	9.1, 9.9	10.1, 10.8	10.3, 10.9	9.3, 9.8
0	10.2, 10.8	10.1, 10.9	11.1, 11.8	11.3, 11.9	9.3, 9.9

a. Identify the experimental design and write an appropriate statistical model.

b. Is there a significant difference in the mean heat losses of the five pane designs? Use $\alpha = .05$.

c. Are the differences in the five designs consistent across the five temperatures? Use $\alpha = .05$ and a profile plot in reaching your conclusion.

d. Use Tukey's W procedure at an $\alpha = .05$ level to compare the mean heat losses for the five pane designs.

Psy. **14.32** An experiment was conducted to examine the effects of different levels of reinforcement and different levels of isolation on children's ability to recall. A single analyst was to work with a random sample of 36 children selected from a relatively homogeneous group of fourth-grade students. Two levels of reinforcement (none and verbal) and three levels of isolation (20, 40, and 60 minutes) were to be used. Students were randomly assigned to the six treatment groups, with a total of six students being assigned to each group.

Each student was to spend a 30-minute session with the analyst. During this time, the student was to memorize a specific passage, with reinforcement provided as dictated by the group to which the student was assigned. Following the 30-minute session, the student was isolated for the time specified for his or her group and then tested for recall of the memorized passage. The data appear next.

Level of Reinforcement	**Time of Isolation (minutes)**					
	20		**40**		**60**	
None	26	19	30	36	6	10
	23	18	25	28	11	14
	28	25	27	24	17	19
Verbal	15	16	24	26	31	38
	24	22	29	27	29	34
	25	21	23	21	35	30

a. What can you conclude about the effects of level of reinforcement and time of isolation on the average recall test score?

b. Verify that the conditions needed to validly apply your tests in part (a) are not violated.

Med. **14.33** Researchers were interested in the stability of a drug product stored for four lengths of time (1, 3, 6, and 9 months). The drug was manufactured with 30 mg/mL of the active ingredient of a drug product, and the amount of the active ingredient in the drug at the end of the storage period was to be determined. The drug was stored at a constant temperature of 30°C. Two laboratories were used in the study, with three 2-mL vials of the drug randomly assigned to each of the four storage times. At the end of the storage time, the amount of the active ingredient was determined for each of the vials. A measure of the pH of the drug was also recorded for each vial. The data are given here.

Time (in months at 30°C)	Laboratory	mg/mL of Active Ingredient	pH	Time (in months at 30°C)	Laboratory	mg/mL of Active Ingredient	pH
1	1	30.03	3.61	1	2	30.12	3.87
1	1	30.10	3.60	1	2	30.10	3.80
1	1	30.14	3.57	1	2	30.02	3.84
3	1	30.10	3.50	3	2	29.90	3.70
3	1	30.18	3.45	3	2	29.95	3.80
3	1	30.23	3.48	3	2	29.85	3.75
6	1	30.03	3.56	6	2	29.75	3.90
6	1	30.03	3.74	6	2	29.85	3.90
6	1	29.96	3.81	6	2	29.80	3.90
9	1	29.81	3.60	9	2	29.75	3.77
9	1	29.79	3.55	9	2	29.85	3.74
9	1	29.82	3.59	9	2	29.80	3.76

a. Write a model relating the pH measured on each vial to the factors of length of storage time and laboratory.

b. Display an analysis of variance table for the model of part (a).

14.34 Refer to Exercise 14.33. Obtain an analysis of variance for both dependent variables (i.e., y_1 = mg/mL of active ingredient and y_2 = pH). Draw conclusions about the stability of these 2-mL vials based on these analyses. Use $\alpha = .05$.

Bus. **14.35** A manufacturer whose daily supply of raw materials is variable and limited can use the material to produce two different products in various proportions. The profit per unit of raw material obtained by producing each of the two products depends on the length of a product's manufacturing run and hence on the amount of raw material assigned to it. Other factors—such as worker productivity, machine breakdown, and so on—can affect the profit per unit as well, but their net effect on profit is random and uncontrollable. The manufacturer has conducted an experiment to investigate the effects of the level of supply of raw material, S, and the ratio of its assignment, R, to the two product manufacturing lines on the profit per unit of raw material. The ultimate goal is to be able to choose the best ratio, R, to match each day's supply of raw materials, S. The levels of supply of the raw material chosen for the experiment were 15, 18, and 21 tons. The levels of the ratio of allocation to the two product lines were 1/2, 1, and 2. The response was the profit (in cents) per unit of raw material supply obtained from a single day's production. Three replications of each combination were conducted in a random sequence. The data for the 27 days are shown in the following table.

Ratio of Raw Material Allocation (R)	Raw Material Supply (tons)		
	15	18	21
1/2	22, 20, 21	21, 19, 20	19, 18, 20
1	21, 20, 19	23, 24, 22	20, 19, 21
2	17, 18, 16	21, 11, 20	20, 22, 24

a. Draw conclusions from an analysis of variance table. Use $\alpha = .05$.
b. Identify the two best combinations of R and S. Are these two combinations significantly different? Use a procedure that limits the error rate of all pairwise comparisons of combinations to be no more than 0.05.

Ag. **14.36** A horticulturalist at a large research institution designs a study to evaluate the effect on tomato yields of water loss due to transpiration. She decides to examine four levels of shading of the tomato plants at three stages of the tomato plant's development. The four levels of shading (0, 25%, 50%, and 75%) were selected to reduce the solar exposure of the plants. The shading remained in place for 20 days during the early, middle, and late phases of the tomato plants' growth. There were four plots of tomatoes randomly assigned to each of the combinations of shading and growth stage. At the end of the study, the yields per plot in pounds were recorded. However, due to a problem in the harvesting of the tomatoes, a few of the plot yields were not recorded.

Growth Stage	Percent Shading			
	0	25%	50%	75%
Early	70.6	57.2	69.5	57.2
	56.3	53.2	55.4	62.9
		44.2		59.0
	55.1	36.7		40.8
Middle	50.5		42.3	78.3
	50.1	67.1	66.0	62.6
	52.7		57.1	58.5
	60.0	62.4	42.5	
Late	69.1	56.8	57.3	61.3
	55.8		67.4	73.3
	43.5	62.1	72.8	
	75.3	75.0	63.0	57.2

a. Identify the design for this experiment.
b. Construct an AOV table for the experiment, and test for the main effects of shading and growth stage and an interaction between shading and growth stage.
c. Is there a linear trend in the mean yields across the levels of percent shading?
d. Which level of shading would you recommend for maximum yield?
e. During which growth stage would you apply the shading?

Env. **14.37** Refer to Exercise 14.36.
a. Are the computational formulas for obtaining the sum of squares appropriate for the data in the tomato experiment? Justify your answer.
b. Verify that there are no major violations in the conditions necessary to conduct the F tests in the AOV table.
c. Write a linear model for this experiment, and estimate all the terms in your model using the data in Exercise 14.36.

Env. **14.38** ***The following experiment is from Kuehl (2000).*** Sludge is a dried product remaining from processed sewage; it contains nutrients beneficial to plant growth. It can be used for fertilizer on agricultural crops provided it does not contain toxic levels of certain elements such as heavy metals (such as zinc, not rock groups). Typically, the levels of metals in sludge are assayed by growing plants in media containing different doses of the sludge.

A soil scientist hypothesized the concentration of certain heavy metals in sludge would differ among the metropolitan areas from which the sludge was obtained. The variation could result from any number of reasons, including the different industrial bases surrounding the areas and the efficiency of the various sewage treatment facilities. If this was true, then recommendations for applications on crops would have to be preceded by knowledge about the source of the sludge material. An assay was planned to determine whether there was significant variation in heavy metal concentrations among diverse metropolitan areas.

The investigator obtained sewage sludge from treatment plants located in three different metropolitan areas. Barley plants were grown in a sand medium to which sludge was added as fertilizer. The sludge was added to the sand at three different rates: 0.5, 1.0, and 1.5 metric tons/acre. Each of the nine treatment combinations was randomly assigned to four replicate containers. The containers were arranged completely at random in a growth chamber. At a certain stage of growth, the zinc contents in parts per million were determined for the barley plants grown in each of the containers. The data are given below.

City A Sludge Rate			City B Sludge Rate			City C Sludge Rate		
0.5	**1.0**	**1.5**	**0.5**	**1.0**	**1.5**	**0.5**	**1.0**	**1.5**
26.4	25.2	26.0	30.1	47.7	73.8	19.4	23.2	18.9
23.5	39.2	44.6	31.0	39.1	71.1	19.3	21.3	19.8
25.4	25.5	35.5	30.8	55.3	68.4	18.7	23.2	19.6
22.9	31.9	38.6	32.8	50.7	77.1	19.0	19.9	21.9

a. Identify the design for this experiment.
b. Write a model for this study. Identify all the terms in your model and any conditions that are placed on the terms.
c. Display estimates of all the parameters in your model.

Env. **14.39** Refer to Exercise 14.38.
a. Construct an AOV table for the experiment, and test for the main effects of sludge rate and source of sludge and an interaction between sludge rate and source of sludge.
b. Is there a linear trend in the mean yields across the sludge rates?
c. Which pairs of sludge rates have significant differences in their mean zinc contents?

Env. **14.40** Refer to Exercise 14.38. Verify that there are no major violations in the conditions necessary to conduct the tests in the AOV table.

CHAPTER 15

Analysis of Variance for Blocked Designs

15.1 Introduction and Abstract of Research Study

In this chapter, we will discuss some standard experimental designs and their analyses. Sections 15.2 and 15.3 introduce extensions of the completely randomized design, where the focus remains the same—namely, treatment mean comparisons—but where other "nuisance" variables must be controlled. In Section 15.4, we discuss designs that combine the attributes of the "block" designs of Sections 15.2 and 15.3 with a factorial treatment structure. The remaining sections of the chapter deal with procedures to check the validity of model conditions and alternative procedures to use when the standard model conditions are not satisfied.

Abstract of Research Study: Control of Leatherjackets

Lawns develop yellow patches during the spring and summer months when the grass has died as a result of leatherjackets (*Tipula* species) eating the roots. Adult leatherjackets of the species (also known as grubs) that damage lawns mainly emerge in late summer and early autumn. The females deposit eggs in the turf and these hatch in the autumn and begin feeding on grass roots. In cold winters, little feeding or development takes place, so signs of damage may not be seen until the summer. However, mild winters can allow the grubs to develop over the winter and sometimes cause damage in late winter or early spring. The larvae have no legs or obvious head, and they have a tough, leathery outer skin. Leatherjackets complete their feeding during the summer and pupate in the soil. Before the adult fly emerges, the pupa wriggles half out of the soil, so the brown pupal case is left sticking out of the turf.

An experiment (designed to evaluate methods for dealing with leather-jackets) is described in the book *A Handbook of Small Data Sets* (Hand et al., 1993). It

TABLE 15.1
Leatherjacket counts on test sites

Plot	Control	Treatment 1	2	3	4
1	33 30	8	12	6	17
	59 36	11	17	10	8
2	36 23	15	6	4	3
	24 23	20	40	7	2
3	19 42	10	12	4	6
	27 39	7	10	12	3
4	71 39	17	5	5	1
	49 20	26	8	5	1
5	22 42	14	12	2	2
	27 22	11	12	6	5
6	84 23	22	16	17	6
	50 37	30	4	11	5

involved a control and four potential chemicals to eliminate the leatherjackets. The data are presented in Table 15.1, and their analysis will be given in Section 15.6.

15.2 Randomized Complete Block Design

In Example 14.1, the researcher was investigating four types of reflective paint used to mark the lanes on rural highways. The paints were applied to sections of highway 6 feet in length. Six months after application of the paint, the percentage decrease in reflectivity was recorded for each of the sections. In this experiment, the researcher had 16 sections of highway for use in the study. The sections were all in the same general location. This type of design did not allow for varying levels of road usage, weather conditions, and maintenance. A new study has been proposed, and the researcher wants to incorporate four different locations into the design of the new study. The researcher identifies 4 sections of roadway 6 feet in length at each of the four locations. If we randomly assigned the four paints to the 16 sections, we might end up with a randomization scheme like the one listed in Table 15.2.

Even though we still have four observations for each treatment in this design, any differences that we may observe among the reflectivities of the road markings for the four types of paint may be due entirely to differences in the road conditions and traffic volumes among the four locations. Because the factors location and type of paint are **confounded,** we cannot determine whether any observed differences in the decrease in reflectivity of the road markings are due to differences in the

confounded

TABLE 15.2
Random assignment of the four paints to the 16 sections

Location 1	2	3	4
P_1	P_2	P_3	P_4
P_1	P_2	P_3	P_4
P_1	P_2	P_3	P_4
P_1	P_2	P_3	P_4

TABLE 15.3
Randomized complete block assignment of the four paints to the 16 sections

Location			
1	**2**	**3**	**4**
P_2	P_2	P_1	P_1
P_1	P_4	P_3	P_2
P_3	P_1	P_4	P_4
P_4	P_3	P_2	P_3

locations of the markings or due to differences in the types of paint used in creating the markings. This example illustrates a situation in which the 16 road markings are affected by an extraneous source of variability: the location of the road markings. If the four locations present different environmental conditions or different traffic volumes, the 16 experimental units would not be a homogeneous set of units on which we could base an evaluation of the effects of the four treatments, the four types of paint.

The completely randomized design just described is not appropriate for this experimental setting. We need to use a randomized complete block design in order to take into account the differences that exist in the experimental units prior to assigning the treatments. In Chapter 2, we described how we can restrict the randomization of treatments to experimental units in order to reduce the variability between experimental units receiving the same treatments. This methodology can be used to ensure that each location has a section of roadway painted with each of the four types of paint. One such randomization is listed in Table 15.3. Note that each location contains four sections of roadway, each section treated with one of the four paints. Hence, the variability in the reflectivity of paints due to differences in roadway conditions at the four locations can now be addressed and controlled. This will allow pairwise comparisons among the four paints that utilize the sample means to be free of the variability among locations. For example, if we ran the test

$$H_0\colon\ \mu_{P_1} - \mu_{P_2} = 0 \ \text{ versus } \ H_a\colon\ \mu_{P_1} - \mu_{P_2} \neq 0$$

and rejected H_0, the differences between μ_{P_1} and μ_{P_2} would be due to a difference between the reflectivity properties of the two paints and not due to a difference among the locations, since both paint P_1 and paint P_2 were applied to a section of roadway at each of the four locations.

In a randomized complete block design, the random assignment of the treatments to the experimental units is conducted separately within each block—the location of the roadways in this example. The four sections within a given location would tend to be more alike with respect to environmental conditions and traffic volume than sections of roadway in two different locations. Thus, we are in essence conducting four independent completely randomized designs, one for each of the four locations. By using the randomized complete block design, we have effectively filtered out the variability among the locations, enabling us to make more precise comparisons among the treatment means μ_{P_1}, μ_{P_2}, μ_{P_3}, and μ_{P_4}.

In general, we can use a randomized complete block design to compare t treatment means when an extraneous source of variability (blocks) is present. If there are b different blocks, we would randomly assign each of the t treatments to an experimental unit in each block in order to filter out the block-to-block variability. In our example, we had $t = 4$ treatments (types of paint) and $b = 4$ blocks (locations).

We can formerly define a randomized complete block design as follows.

DEFINITION 15.1

A **randomized complete block design** is an experimental design for comparing t treatments in b blocks. The blocks consist of t homogeneous experimental units. Treatments are randomly assigned to experimental units within a block, with each treatment appearing exactly once in every block.

The randomized complete block design has certain advantages and disadvantages, as shown here.

Advantages and Disadvantages of the Randomized Complete Block Design

Advantages

1. The design is useful for comparing t treatment means in the presence of a single extraneous source of variability.
2. The statistical analysis is simple.
3. The design is easy to construct.
4. The design can be used to accommodate any number of treatments in any number of blocks.

Disadvantages

1. Because the experimental units within a block must be homogeneous, the design is best suited for a relatively small number of treatments.
2. This design controls for only one extraneous source of variability (due to blocks). Additional extraneous sources of variability tend to increase the error term, making it more difficult to detect treatment differences.
3. The effect of each treatment on the response must be approximately the same from block to block.

Consider the data for a randomized complete block design as arranged in Table 15.4. Note that although these data look similar to the data presentation for a completely randomized design (see Table 14.2), there is a difference in the way treatments were assigned to the experimental units.

TABLE 15.4
Data for a randomized complete block design

	Block				
Treatment	**1**	**2**	**⋯**	***b***	**Mean**
1	y_{11}	y_{12}	⋯	y_{1b}	$\bar{y}_{1.}$
2	y_{21}	y_{22}	⋯	y_{2b}	$\bar{y}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
t	y_{t1}	y_{t2}	⋯	y_{tb}	$\bar{y}_{t.}$
Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	⋯	$\bar{y}_{.b}$	$\bar{y}_{..}$

The model for an observation in a randomized complete block design can be written in the form

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

where the terms of the model are defined as follows:

y_{ij}: Observation on experimental unit in jth block receiving treatment i.

μ: Overall mean, an unknown constant.

τ_i: An effect due to treatment i, an unknown constant.

β_j: An effect due to block j, an unknown constant.

ε_{ij}: A random error associated with the response from an experimental unit in block j receiving treatment i. We require that the ε_{ij}s have a normal distribution with mean 0 and common variance σ_ε^2. In addition, the errors must be independent.

The conditions given above for our model can be shown to imply that the recorded response from the ith treatment in the jth block, y_{ij}, is normally distributed with mean

$$\mu_{ij} = E(y_{ij}) = \mu + \tau_i + \beta_j$$

and variance σ_ε^2. Table 15.5 gives the population means (expected values) for the data of Table 15.4.

Similarly to the model for a completely randomized design, the above model is overparametrized. In order to obtain the least-squares estimators, we need to place the following constraints on the effect parameters: $\tau_t = 0$ and $\beta_b = 0$.

Under the above constraints, the relationship among the parameters μ, τ_i, and β_i and the treatment means, $\mu_{ij} = \mu + \tau_i + \beta_j$, becomes

a. Overall mean: $\mu = \mu_{tb}$
b. Main effects of factor A: $\tau_i = \mu_{ib} - \mu_{tb}$ for $i = 1, 2, \ldots, t-1$
c. Main effects of blocks: $\beta_j = \mu_{tj} - \mu_{tb}$ for $j = 1, 2, \ldots, b-1$

Several comments should be made concerning the table of expected values. First, any pair of observations that receive the same treatment (appear in the same row of Table 15.5) has population means that differ only by their block effects (β_js). For example, the expected values associated with y_{11} and y_{12} (two observations receiving treatment 1) are

$$\mu_{11} = \mu + \tau_1 + \beta_1 \qquad \mu_{12} = \mu + \tau_1 + \beta_2$$

Thus, the difference in their means is

$$\mu_{11} - \mu_{12} = (\mu + \tau_1 + \beta_1) - (\mu + \tau_1 + \beta_2) = \beta_1 - \beta_2$$

TABLE 15.5
Expected values for the y_{ij}s in a randomized block design

	Block			
Treatment	**1**	**2**	**⋯**	***b***
1	$\mu_{11} = \mu + \tau_1 + \beta_1$	$\mu_{12} = \mu + \tau_1 + \beta_2$	⋯	$\mu_{1b} = \mu + \tau_1 + \beta_b$
2	$\mu_{21} = \mu + \tau_2 + \beta_1$	$\mu_{22} = \mu + \tau_2 + \beta_2$	⋯	$\mu_{2b} = \mu + \tau_2 + \beta_b$
⋮	⋮	⋮	⋮	⋮
t	$\mu_{t1} = \mu + \tau_t + \beta_1$	$\mu_{t2} = \mu + \tau_t + \beta_2$	⋯	$\mu_{tb} = \mu + \tau_t + \beta_b$

which accounts for the fact that y_{11} was recorded in block 1 and y_{12} was recorded in block 2 but both were responses from experimental units receiving treatment 1. Thus, there is no treatment effect, but a block effect may be present. Second, two observations appearing in the same block (in the same column of Table 15.5) have means that differ by a treatment effect only. For example, y_{11} and y_{21} both appear in block 1. The difference in their means, from Table 15.5, is

$$\mu_{11} - \mu_{21} = (\mu + \tau_1 + \beta_1) - (\mu + \tau_2 + \beta_1) = \tau_1 - \tau_2$$

which accounts for the fact that the experimental units received different treatments but were observed in the same block. Hence, there may be a treatment effect but no block effect. Finally, when two experimental units receive different treatments and are observed in different blocks, their expected values differ by effects due to both treatment differences and block differences. Thus, observations y_{11} and y_{22} have expectations that differ by

$$\mu_{11} - \mu_{22} = (\mu + \tau_1 + \beta_1) - (\mu + \tau_2 + \beta_2) = (\tau_1 - \tau_2) + (\beta_1 - \beta_2)$$

filtering

Using the information we have learned concerning the model for a randomized block design, we can illustrate the concept of **filtering** and show how the randomized block design filters out the variability due to blocks. Consider a randomized block design with $t = 3$ treatments (1, 2, and 3) laid out in $b = 3$, blocks, as shown in Table 15.6.

The model for this randomized block design is

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \quad (i = 1, 2, 3; j = 1, 2, 3)$$

Suppose we wish to estimate the difference in mean responses for treatments 2 and 1—namely, $\mu_{2.} - \mu_{1.}$. The difference in sample means, $\bar{y}_{2.} - \bar{y}_{1.}$, would represent a point estimate of $\mu_{2.} - \mu_{1.}$. By substituting into our model, we have

$$\begin{aligned}\bar{y}_{1.} &= \frac{1}{3}\sum_j y_{1j} \\ &= \frac{1}{3}[(\mu + \tau_1 + \beta_1 + \varepsilon_{11}) + (\mu + \tau_1 + \beta_2 + \varepsilon_{12}) + (\mu + \tau_1 + \beta_3 + \varepsilon_{13})] \\ &= \mu + \tau_1 + \bar{\beta}_{.} + \bar{\varepsilon}_{1.}\end{aligned}$$

where $\bar{\beta}_{.}$ represents the mean of the three block effects—β_1, β_2, and β_3—and $\bar{\varepsilon}_{1.}$ represents the mean of the three random errors—ε_{11}, ε_{12}, and ε_{13}. Similarly, it is easy to show that

$$\bar{y}_{2.} = \mu + \tau_2 + \bar{\beta}_{.} + \bar{\varepsilon}_{2.}$$

and hence

$$\bar{y}_{2.} - \bar{y}_{1.} = (\tau_2 - \tau_1) + (\bar{\varepsilon}_{2.} - \bar{\varepsilon}_{1.})$$

Note how the block effects cancel, leaving the quantity $(\bar{\varepsilon}_{2.} - \bar{\varepsilon}_{1.})$ as the error of estimation using $\bar{y}_{2.} - \bar{y}_{1.}$ to estimate $(\mu_{2.} - \mu_{1.})$.

TABLE 15.6
Randomized complete block design with $t = 3$ treatments and $b = 3$ blocks

Block	Treatment		
1	1	2	3
2	1	3	2
3	3	1	2

If a completely randomized design had been employed instead of a randomized block design, treatments would have been assigned to experimental units at random, and it is quite likely that a treatment will appear more than once in some block and hence one or more of the treatments will not appear in that block. When the same treatment appears more than once in a block and we calculate an estimate of $(\mu_{2.} - \mu_{1.})$ using $\bar{y}_{2.} - \bar{y}_{1.}$, all block effects would not cancel out as they did previously. Then the error of estimation would include not only $\bar{\varepsilon}_{2.} - \bar{\varepsilon}_{1.}$ but also the block effects that do not cancel; that is,

$$\bar{y}_{2.} - \bar{y}_{1.} = \tau_2 - \tau_1 + [(\bar{\varepsilon}_{2.} - \bar{\varepsilon}_{1.}) + (\text{block effects that do not cancel})]$$

Hence, the randomized block design filters out variability due to blocks by decreasing the error of estimation for a comparison of treatment means.

A plot of the expected values, μ_{ij} in Figure 15.1, demonstrates that the size of the difference between the means of observations receiving the same treatment but in different blocks (say, j and j') is the same for all treatments. That is,

$$\mu_{ij} - \mu_{ij'} = \beta_j - \beta_{j'} \qquad \text{for all } i = 1, \ldots t$$

A consequence of this condition is that the lines connecting the means having the same treatment form a set of parallel lines.

The main goal in using the randomized complete block design was to examine differences in the t treatment means $\mu_{1.}, \mu_{2.}, \ldots, \mu_{t.}$, where μ_i is the mean response of treatment i. The null hypothesis is *no difference among treatment means* versus the research hypothesis, which is *treatment means differ.* That is,

$$H_0\text{: } \mu_{1.} = \mu_{2.} = \cdots = \mu_{t.} \text{ versus } H_a\text{: At least one } \mu_{i.} \text{ differs from the rest.}$$

This set of hypothesis is equivalent to testing

$$H_0\text{: } \tau_1 = \tau_2 = \cdots = \tau_t = 0 \text{ versus } H_a\text{: At least one } \tau_i \text{ differs from 0.}$$

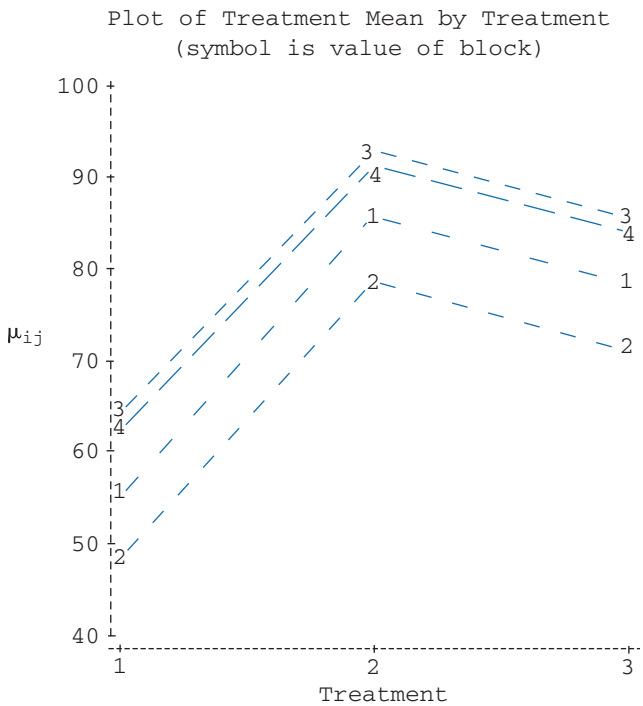


FIGURE 15.1
Treatment means in a randomized block design with four blocks

The two sets of hypotheses are equivalent because, as we observed in Table 15.5, when comparing the mean responses of two treatments (say, i and i') observed in the same block, the difference in their mean responses is

$$\mu_{i.} - \mu_{i'.} = \tau_i - \tau_{i'}$$

Thus, under H_0, we are assuming that treatments have the same mean responses within a given block. Our test statistic will be obtained by examining the model for a randomized block design and partitioning the total sum of squares to include terms for treatment effects, block effects, and random error effects. Using Table 15.4, we can introduce notation that is needed in the partitioning of the total sum of squares. This notation is presented here.

y_{ij}: Observation for treatment i in block j

t: Number of treatments

b: Number of blocks

$\bar{y}_{i.}$: Sample mean for treatment i, $\bar{y}_{i.} = \frac{1}{b}\sum_{j=1}^{b} y_{ij}$

$\bar{y}_{.j}$: Sample mean for block j, $\bar{y}_{.j} = \frac{1}{t}\sum_{i=1}^{t} y_{ij}$

$\bar{y}_{..}$: Overall sample mean, $\bar{y}_{..} = \frac{1}{tb}\sum_{ij} y_{ij}$

total sum of squares

The **total sum of squares** of the measurements about their mean $\bar{y}_{..}$ is defined as before:

$$\text{TSS} = \sum_{ij}(y_{ij} - \bar{y}_{..})^2$$

This sum of squares will be partitioned into three separate sources of variability: one due to the variability among treatments, one due to the variability among blocks, and one due to the variability from all sources not accounted for by either treatment differences or block differences. We call this source of variability **error.**

error

partition of TSS

The **partition of TSS** is similar to the partition from Chapter 14 for a two-factor treatment structure without an interaction term.

It can be shown algebraically that TSS takes the following form:

$$\sum_{ij}(y_{ij} - \bar{y}_{..})^2 = b\sum_{i}(\bar{y}_{i.} - \bar{y}_{..})^2 + t\sum_{j}(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{ij}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

The first quantity on the right-hand side of the equal sign measures the variability of the treatment means $\bar{y}_{i.}$ from the overall mean $\bar{y}_{..}$. Thus,

$$\text{SST} = b\sum_{i}(\bar{y}_{i.} - \bar{y}_{..})^2$$

between-treatment sum of squares

called the **between-treatment sum of squares,** is a measure of the variability in the y_{ij}s due to differences in the treatment means. Similarly, the second quantity,

$$\text{SSB} = t\sum_{j}(\bar{y}_{.j} - \bar{y}_{..})^2$$

between-block sum of squares

measures the variability between the block means $\bar{y}_{.j}$ and the overall mean. It is called the **between-block sum of squares.** The third source of variability, referred

TABLE 15.7
Analysis of variance table for a randomized complete block design

Source	SS	df	MS	*F*
Treatments	SST	$t-1$	$\text{MST} = \text{SST}/(t-1)$	MST/MSE
Blocks	SSB	$b-1$	$\text{MSB} = \text{SSB}/(b-1)$	MSB/MSE
Error	SSE	$(b-1)(t-1)$	$\text{MSE} = \text{SSE}/(b-1)(t-1)$	
Total	TSS	$bt-1$		

sum of squares for error

to as the **sum of squares for error,** SSE, represents the variability in the $\bar{y}_{ij}$s not accounted for by the block and treatment differences. There are several forms for this term:

$$\text{SSE} = \sum_{ij}(e_{ij})^2 = \sum_{ij}(y_{ij} - \bar{y}_{j.} - \bar{y}_{.j} + \bar{y}_{..})^2 = \text{TSS} - \text{SST} - \text{SSB}$$

where $e_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j$ are the residuals used to check model conditions. We can summarize our calculations in an AOV table, as given in Table 15.7.

The hypotheses for testing differences in the treatment means are

$$H_0\text{: } \tau_1 = \tau_2 = \cdots = \tau_t = 0 \text{ versus } H_a\text{: At least one } \tau_i \text{ is different from zero.}$$

In terms of the treatment means, $\mu_{i.}$, H_0 and H_a can be written as

$$H_0\text{: } \mu_{1.} = \mu_{2.} = \cdots = \mu_{t.} \quad H_a\text{: At least one } \mu_{i.} \text{ is different from the rest.}$$

The test statistic for testing these hypotheses is the ratio

$$F = \frac{\text{MST}}{\text{MSE}}$$

unbiased estimates

expected mean squares

When H_0: $\mu_{1.} = \mu_{2.} = \cdots = \mu_{t.}$ is true, both MST and MSE are **unbiased estimates** of σ_ε^2, the variance of the experimental error. That is, when H_0 is true, both MST and MSE have mean values in repeated sampling, called the **expected mean squares,** equal to σ_ε^2. We express these terms as

$$E(\text{MST}) = \sigma_\varepsilon^2 \quad E(\text{MSE}) = \sigma_\varepsilon^2$$

We would thus expect $F = \text{MST/MSE}$ to have a value near 1.

When H_a is true, the expected value of MSE is still σ_ε^2. However, MST is no longer unbiased for σ_ε^2. In fact, the expected mean square for treatments can be shown to be

$$E(\text{MST}) = \sigma_\varepsilon^2 + b\theta_T, \text{ where } \theta_T = \frac{1}{t-1}\sum_{i=1}^{t}(\mu_{i.} - \mu_{..})^2$$

Thus, a large difference in the treatment means will result in a large value for θ_T. The expected value of MST will then be larger than the expected value of MSE, and we would expect $F = \text{MST}/\text{MSE}$ to be larger than 1. Thus, our test statistic F rejects H_0 when we observe a value of F larger than a value in the upper tail of the F distribution.

The above discussion leads to the following decision rule for a specified probability of a Type I error:

$$\text{Reject } H_0\text{: } \mu_{1.} = \mu_{2.} = \cdots = \mu_{t.} \text{ when } F = \text{MST/MSE exceeds } F_{\alpha,\,\text{df}_1,\,\text{df}_2}$$

where $F_{\alpha,\,\text{df}_1,\,\text{df}_2}$ is from the F tables in Appendix Table 8 with α = specified value of probability of Type I error, $\text{df}_1 = \text{df}_{\text{MST}} = t - 1$, and $\text{df}_2 = \text{df}_{\text{MSE}} = (b-1)(t-1)$.

Alternatively, we can compute the p-value for the observed value of the test statistic F_{obs} by computing

$$p\text{-value} = P(F_{\text{df}_1,\,\text{df}_2} > F_{\text{obs}}) = 1\ \textbf{\textit{pf}}(F_{\text{obs}}, t-1, (b-1)(t-1))$$

where the F distribution with $\text{df}_1 = t - 1$ and $\text{df}_2 = (b-1)(t-1)$ is used to compute the probability. We would then compare the p-value to a selected value for the probability of Type I error, with small p-values supporting the research hypothesis and large p-values failing to reject H_0.

The block effects are generally assessed only to determine whether or not the blocking was efficient in reducing the variability in the experimental units. Thus, hypotheses about the block effects are not tested. However, we might still ask whether blocking has increased our precision for comparing treatment means in a given experiment. Let MSE_{RCB} and MSE_{CR} denote the mean square errors for a randomized complete block design and a completely randomized design, respectively. One measure of precision for the two designs is the variance of the estimate of the ith treatment mean, $\hat{\mu}_{i.} = \bar{y}_{i.}$ $(i = 1, 2, \ldots, t)$. For a randomized complete block design, the estimated variance of $\bar{y}_{i.}$ is $\text{MSE}_{\text{RCB}}/b$. For a completely randomized design, the estimated variance of $\bar{y}_{i.}$ is MSE_{CR}/r, where r is the number of observations (replications) of each treatment required to satisfy the relationship

$$\frac{\text{MSE}_{\text{CR}}}{r} = \frac{\text{MSE}_{\text{RCB}}}{b} \quad \text{or} \quad \frac{\text{MSE}_{\text{CR}}}{\text{MSE}_{\text{RCB}}} = \frac{r}{b}$$

relative efficiency RE(RCB, CR)

The quantity r/b is called the **relative efficiency** of the randomized complete block design compared to a completely randomized design **RE(RCB, CR).** The larger the value of MSE_{CR} is compared that of to MSE_{RCB}, the larger r must be to obtain the same level of precision for estimating a treatment mean in a completely randomized design as obtained using the randomized complete block design. Thus, if the blocking is effective, we would expect the variability in the experimental units to be smaller in the randomized complete block design than in a completely randomized design. The ratio $\text{MSE}_{\text{CR}}/\text{MSE}_{\text{RCB}}$ should be large, which would result in r being much larger than b. Thus, the amount of data needed to obtain the same level of precision in estimating μ_i would be larger in the completely randomized design than in the randomized complete block design. When the blocking is not effective, then the ratio $\text{MSE}_{\text{CR}}/\text{MSE}_{\text{RCB}}$ would be nearly 1, and r and b would be equal.

In practice, evaluating the efficiency of the randomized complete block design relative to that of a completely randomized design cannot be accomplished because the completely randomized design was not conducted. However, we can use the mean squares from the randomized complete block design, MSB and MSE, to obtain the relative efficiency RE(RCB, CR) by using the formula

$$\text{RE(RCB, CR)} = \frac{\text{MSE}_{\text{CR}}}{\text{MSE}_{\text{RCB}}} = \frac{(b-1)\text{MSB} + b(t-1)\text{MSE}}{(bt-1)\text{MSE}}$$

When RE(RCB, CR) is much larger than 1, then r is greater than b, and we would conclude that the blocking was efficient because many more observations would be required in a completely randomized design than would be required in the randomized complete block design.

EXAMPLE 15.1

A researcher conducted an experiment to compare the effects of three different insecticides on a variety of string beans. To obtain a sufficient amount of data, it was necessary to use four different plots of land. Since the plots had somewhat different soil fertility, drainage characteristics, and sheltering from winds, the researcher decided to conduct a randomized complete block design with the plots serving as the blocks. Each plot was subdivided into three rows. A suitable distance was maintained between rows within a plot so that the insecticides could be confined to a particular row. Each row was planted with 100 seeds and then maintained under the insecticide assigned to the row. The insecticides were randomly assigned to the rows within a plot so that each insecticide appeared in one row within all four plots. The response y_{ij} of interest was the number of seedlings that emerged per row. The data and means are given in Table 15.8.

TABLE 15.8
Number of seedlings by insecticide and plot for Example 15.1

	Plot				
Insecticide	**1**	**2**	**3**	**4**	**Insecticide Mean**
1	56	48	66	62	58
2	83	78	94	93	87
3	80	72	83	85	80
Plot mean	73	66	81	80	75

a. Write an appropriate statistical model for this experimental situation.
b. Run an analysis of variance to compare the effectiveness of the three insecticides. Use $\alpha = .05$.
c. Summarize your results in an AOV table.
d. Compute the relative efficiency of the randomized block design relative to a completely randomized design.

Solution We recognize this experimental design as a randomized complete block design with $b = 4$ blocks (plots) and $t = 3$ treatments (insecticides) per block. The appropriate statistical model is

a. $y_{ij} = \mu + t_i + \beta_j + \varepsilon_{ij} \quad i = 1, 2, 3;\ j = 1, 2, 3, 4$

From the information in Table 15.8, we can estimate the treatment means, $\mu_{i.}$, by $\hat{\mu}_{i.} = \bar{y}_{i.}$, which yields

$$\hat{\mu}_{1.} = 58 \quad \hat{\mu}_{2.} = 87 \quad \hat{\mu}_{3.} = 80$$

It would appear that the rows treated with insecticide 1 yielded many fewer plants than the other two insecticides. We will next construct the AOV table.

b. Substituting into the formulas for the sum of squares, we have

$$\text{TSS} = \sum_{ij}(y_{ij} - \bar{y}_{..})^2 = (56 - 75)^2 + (48 - 75)^2 + \cdots + (85 - 75)^2 = 2{,}296$$

$$\text{SST} = b\sum_{i}(\bar{y}_{i.} - \bar{y}_{..})^2 = 4[(58 - 75)^2 + (87 - 75)^2 + (80 - 75)^2] = 1{,}832$$

$$\text{SSB} = t\sum_{j}(\bar{y}_{j} - \bar{y}_{..})^2 = 3[(73 - 75)^2 + (66 - 75)^2 + (81 - 75)^2 + (80 - 75)^2]$$
$$= 438$$

By subtraction, we have

$$\text{SSE} = \text{TSS} - \text{SST} - \text{SSB} = 2{,}296 - 1{,}832 - 438 = 26$$

The analysis of variance table in Table 15.9 summarizes our results. Note that the mean square for a source in the AOV table is computed by dividing the sum of squares for that source by its degrees of freedom.

c.

TABLE 15.9
AOV table for the data of Example 15.1

Source	SS	df	MS	F	p-value
Treatments	1,832	2	916	211.38	.0001
Blocks	438	3	146	33.69	.0004
Error	26	6	4.3333		
Total	2,296	11			

The F test for differences in the treatment means

H_0: $\mu_{1.} = \mu_{2.} = \mu_{3.}$ versus H_a: At least one $\mu_{i.}$ is different from the rest.

Makes use of the F statistic MST/MSE. Since the computed value of F, 211.38, is greater than the tabulated F-value, 5.14, based on $df_1 = 2$, $df_2 = 6$, and $\alpha = .05$, we reject H_0 and conclude that there is significant evidence (p-value $= 1 - \boldsymbol{pf}(211.38, 2, 6) = .0000027$) of a difference in the mean number of seedlings among the three insecticides.

d. We will next assess whether the blocking was effective in increasing the precision of the analysis relative to a completely randomized design. From the AOV table, we have MSB = 146 and MSE = 4.3333. Hence, the relative efficiency of this randomized block design relative to a completely randomized design is

$$\text{RE(RCB, CR)} = \frac{(b-1)\text{MSB} + b(t-1)\text{MSE}}{(bt-1)\text{MSE}}$$

$$= \frac{(4-1)(146) + 4(3-1)(4.3333)}{[(4)(3)-1](4.3333)} = 9.92$$

That is, approximately 10 times as many observations of each treatment would be required in a completely randomized design to obtain the same precision for estimating the treatment means as with this randomized complete block design. The plots were considerably different in their physical characteristics, and, hence, it was crucial that blocking be used in this experiment. ■

The results in Example 15.1 are valid only if we can be assured that the conditions placed on the model are consistent with the observed data. Thus, we use the residuals $e_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j$ to assess whether the conditions of normality, equal variance, and independence appear to be satisfied for the observed data. The following example includes the computer output for such an analysis.

EXAMPLE 15.2

The computer output for the experiment described in Example 15.1 is displayed here. Compare the results to those obtained using the definition of the sum of squares, and assess whether the model conditions appear to be valid.

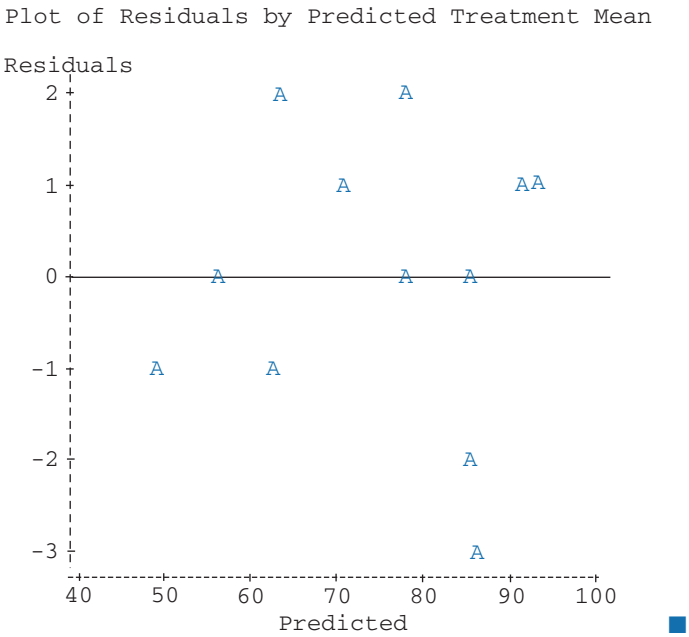
```
Dependent Variable: NUMBER OF SEEDLINGS

                                      Sum of          Mean
Source                        DF      Squares        Square    F Value    Pr > F
Model                          5    2270.0000      454.0000     104.77    0.0001
Error                          6      26.0000        4.3333
Corrected Total               11    2296.0000

Source                        DF    Type I SS   Mean Square    F Value    Pr > F
INSECTICIDES                   2    1832.0000      916.0000     211.38    0.0001
PLOTS                          3     438.000       146.0000      33.69    0.0004

RESIDUAL ANALYSIS

Variable=RESIDUALS

                                Moments

          N                        12  Sum Wgts           12
          Mean                      0  Sum                 0
          Std Dev            1.537412  Variance      2.363636
          Skewness          -0.54037   Kurtosis     -0.25385
          W:Normal          0.942499   Pr<W           0.4938   <- Test of Normality

        Stem Leaf                     #             Boxplot
           2 00                       2                |
           1 000                      3             +-----+
           0 000                      3             *--+--*
          -0                                        |     |
          -1 00                       2             +-----+
          -2 0                        1                |
          -3 0                        1                |
             ----+----+----+----+
```

```
Variable=RESIDUALS

                        Normal Probability Plot
 2.5+                                                  *     ++*+++
    |                                          * * +*+++++
    |                                  * * +*+++++
-0.5+                          * +*++++
    |                    ++*++++
    |          +++*++
-3.5+++++++
     +----+----+----+----+----+----+----+----+----+----+
          -2         -1          0         +1         +2
```

Solution Note that our hand calculations yielded the same values as are given in the computer output. Generally, there will be some rounding errors in our hand calculations, which can lead to values that will differ from those given in the computer output. It is strongly recommended that a computer software program be used in the analysis of variance calculations because of the potential for rounding errors. In assessing whether the model conditions have been met, we first note that in regard to the normality condition, the test of H_0: residuals have normal distribution; the p-value from the Shapiro–Wilks test is p-value = .4938. Thus, we would not reject H_0, and the normality condition appears to be satisfied. Also, the stem and leaf plot, boxplot, and normal probability plot are consistent with the condition

that the residuals have a normal distribution. Figure 15.2 is a plot of the residuals versus the estimated treatment means. From this plot, it would appear that the variability in the residuals is somewhat constant across the treatments.

FIGURE 15.2
Residuals versus treatment means from Example 15.1



15.3 Latin Square Design

Latin square design

The randomized complete block design is used when there is one factor of interest and the experimenter wants to control a single source of extraneous variation. When there are two possible sources of extraneous variation, a **Latin square design** is the appropriate design for the experiment. Consider the following example.

EXAMPLE 15.3

A nonprofit consumer-product testing organization is in the process of evaluating five major brands of room air cleaners. In order to make the ratings as realistic as possible, the organization's engineers decided to evaluate the air cleaners outside the testing laboratory in residential homes. To control for variations due to the differing air qualities in the homes and due to the time-of-the-year characteristics of external air pollution, the engineers decided to use a cleaner of each brand in each of five homes and to run the tests at five different months. The factors to be considered in the study are

1. Brand of air cleaner: B_1, B_2, B_3, B_4, B_5
2. Residential home: H_1, H_2, H_3, H_4, H_5
3. Month: M_1, M_2, M_3, M_4, M_5

The two factors, home and month of the year, are extraneous sources of variation that are important to include in the study in order to provide a more precise evaluation of the differences in the five brands. However, these factors are not of

central importance to the engineers. The response variable is the clean air delivery rate (CADR). CADR is a measure of the air cleaner's ability to reduce smoke, dust, and pollen particles from the air. CADR is defined as the rate of contaminant reduction in the room when the air cleaner is turned on, minus the rate of natural decay when the unit is not running, multiplied by the volume of air in the room, measured in cubic feet. The engineers initially considered using the completely randomized block design displayed in Table 15.10, with brands as treatments and homes as blocks.

TABLE 15.10
A randomized complete block design for the air cleaner study

	Home				
Month	**1**	**2**	**3**	**4**	**5**
M_1	B_2	B_2	B_3	B_2	B_2
M_2	B_1	B_4	B_4	B_4	B_5
M_3	B_3	B_1	B_2	B_5	B_4
M_4	B_5	B_3	B_1	B_3	B_1
M_5	B_4	B_5	B_5	B_1	B_3

In this design, the brand of air cleaner is randomly assigned to the month separately for each of the five homes. Suppose the time of the year, month, has an important impact on the performance of the air cleaner. In the spring, the pollen count may be very high in some areas of the country, or because of wind patterns, industrial air pollution could be considerably higher during some months and very low during other months. The design in Table 15.10 would then produce a strong positive bias for brand B_2 if month M_1 had the lowest levels of air particles relative to the other four months because B_2 was observed four times in this month. Similarly, brand B_4 would have a strong negative bias if month M_2 had higher levels of air particles relative to the other four months. Thus, if it is found that brand B_2 produced the highest average CADR, the organization could not be certain whether the brand B_2 was the better air cleaner or whether the results were due to having four of the five tests run during a month in which the air particle level was very low.

This example illustrates a situation in which the experimental units (rooms in home) are affected by two sources of extraneous variation, the home and the month of the year. We can modify the randomized complete block design to filter out this second source of variability, the variability among months, in addition to filtering out the first source, variability among homes. To do this, we restrict our randomization to ensure that each treatment appears in each row (month) and in each column (home). One such randomization is shown in Table 15.11. Note that

TABLE 15.11
A Latin square design for the air cleaner study

	Home				
Month	**1**	**2**	**3**	**4**	**5**
M_1	B_1	B_2	B_3	B_4	B_5
M_2	B_2	B_3	B_4	B_5	B_1
M_3	B_3	B_4	B_5	B_1	B_2
M_4	B_4	B_5	B_1	B_2	B_3
M_5	B_5	B_1	B_2	B_3	B_4

the brands of air cleaners have been assigned to month and home so that each brand is evaluated once in each of the months and homes. Hence, pairwise comparisons among brands that involve the sample means have been adjusted for the variability among months and homes. ■

Latin square design

This experimental design is called a **Latin square design.** In general, a Latin square design can be used to compare t treatment means in the presence of two extraneous sources of variability, which we block off into t rows and t columns. The t treatments are then randomly assigned to the rows and columns so that each treatment appears in every row and every column of the design (see Table 15.11).

The advantages and disadvantages of the Latin square design are listed here.

Advantages and Disadvantages of the Latin Square Design

Advantages

1. The design is particularly appropriate for comparing t treatment means in the presence of two sources of extraneous variation, each measured at t levels.
2. The analysis is quite simple.
3. A Latin square can be constructed for any value of t.

Disadvantages

1. Any additional extraneous sources of variability tend to inflate the error term, making it more difficult to detect differences among the treatment means.
2. The effect of each treatment on the response must be approximately the same across rows and columns.

The definition of a Latin square design is given here.

DEFINITION 15.2

A $t \times t$ **Latin square design** contains t rows and t columns. The t treatments are randomly assigned to experimental units within the rows and columns so that each treatment appears in every row and in every column.

The model for a response in a Latin square design can be written in the form

$$y_{ijk} = \mu + \tau_k + \beta_i + \gamma_j + \varepsilon_{ijk}$$

where the terms of the model are defined as follows:

- y_{ijk}: Observation on experimental unit in the ith row and jth column receiving treatment k.
- μ: Overall mean, an unknown constant.
- τ_k: An effect due to treatment k, an unknown constant.
- β_i: An effect due to row i, an unknown constant.
- γ_j: An effect due to column, j, an unknown constant.

e_{ijk}: A random error associated with the response from an experimental unit in row i and column j. We require that the ε_{ijk}s have a normal distribution with mean 0 and common variance σ_ε^2. In addition, the errors must be independent.

The conditions given above for our model can be shown to imply that the recorded response in the ith row and jth column, y_{ijk}, is normally distributed with mean

$$\mu_{ijk} = E(y_{ijk}) = \mu + \tau_k + \beta_i + \gamma_j$$

additive

and variance σ_ε^2. This model is a completely **additive** model in that there are no interaction terms. The row-blocking variable and column-blocking variable do not interact with the treatment or with each other. Because we have only one observation in each of the cells, only two of the three subscripts on y_{ijk} are necessary to denote a particular response. For example, in Table 15.11 for the response in row 2 and column 4, we have $i = 2$ and $j = 4$; then we automatically know that brand B_5 was used—that is, $k = 5$. This result occurs because each treatment appears exactly once in each row and in each column.

filtering

We can use the model to illustrate how a Latin square design **filters** out extraneous variability due to row and column sources of variability. Here we will consider a Latin square design with $t = 4$ treatments (I, II, III, and IV) and two sources of extraneous variability, each with $t = 4$ levels. This design is displayed in Table 15.12.

If we wish to estimate $\mu_{..3} - \mu_{..1}$, the difference in the mean responses for treatments III and I, using the difference in sample means $\bar{y}_{..3} - \bar{y}_{..1}$, we can substitute into our model to obtain expressions for $\bar{y}_{..3}$ and $\bar{y}_{..1}$, carefully noting in which rows and columns the treatments appear. With y_{ijk} denoting the observation in row i and column j, we have, from Table 15.12,

$$\begin{aligned}\bar{y}_{..1} &= \frac{1}{4}(y_{111} + y_{241} + y_{331} + y_{421}) \\ &= \mu + \tau_1 + \frac{1}{4}(\beta_1 + \beta_2 + \beta_3 + \beta_4) + \frac{1}{4}(\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4) + \bar{\varepsilon}_{..1}\end{aligned}$$

where $\bar{\varepsilon}_{..1}$ is the mean of the random errors for the four observations on treatment I. Similarly,

$$\begin{aligned}\bar{y}_{..3} &= \frac{1}{4}(y_{133} + y_{223} + y_{313} + y_{443}) \\ &= \mu + \tau_3 + \frac{1}{4}(\beta_1 + \beta_2 + \beta_3 + \beta_4) + \frac{1}{4}(\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4) + \bar{\varepsilon}_{..3}\end{aligned}$$

TABLE 15.12
A 4×4 Latin square design

	Column			
Row	**1**	**2**	**3**	**4**
1	I	II	III	IV
2	II	III	IV	I
3	III	IV	I	II
4	IV	I	II	III

Then the sample difference is

$$\bar{y}_{\cdot\cdot 3} - \bar{y}_{\cdot\cdot 1} = \tau_3 - \tau_1 + (\bar{\varepsilon}_{\cdot\cdot 3} - \bar{\varepsilon}_{\cdot\cdot 1})$$

and the error of estimation for $\tau_3 - \tau_1$ is $\bar{\varepsilon}_{\cdot\cdot 3} - \bar{\varepsilon}_{\cdot\cdot 1}$.

If a randomized block design had been used with blocks representing columns, treatments would be randomized within the columns only. It is quite possible for the same treatment to appear more than once in the same row. Then the sample difference would be

$$\bar{y}_{\cdot\cdot 3} - \bar{y}_{\cdot\cdot 1} = \tau_3 - \tau_1 + [(\bar{\varepsilon}_{\cdot\cdot 3} - \bar{\varepsilon}_{\cdot\cdot 1}) + (\text{row effects that do not cancel})]$$

Thus, the error of estimation would be inflated by the row effects that do not cancel out.

EXAMPLE 15.4

Suppose the design displayed in Table 15.12 would have been run as a randomized block design with the four treatments randomly assigned to the rows within each column. One possible randomization is presented in Table 15.13. Show that the difference in the sample means for treatments I and III involves row effects and hence that treatment effects are confounded with row effects.

TABLE 15.13
A randomized block design for four treatments (columns are blocks)

	Column			
Row	**1**	**2**	**3**	**4**
1	II	II	III	II
2	I	IV	IV	IV
3	III	I	II	I
4	IV	III	I	III

Solution We first compute

$$\begin{aligned}\bar{y}_{\cdot\cdot 3} &= \frac{1}{4}(y_{133} + y_{313} + y_{423} + y_{443}) \\ &= \mu + \tau_3 + \frac{1}{4}(\beta_1 + \beta_3 + \beta_4 + \beta_4) + \frac{1}{4}(\gamma_3 + \gamma_1 + \gamma_2 + \gamma_4) + \bar{\varepsilon}_{\cdot\cdot 3} \\ \bar{y}_{\cdot\cdot 1} &= \frac{1}{4}(y_{211} + y_{321} + y_{341} + y_{431}) \\ &= \mu + \tau_1 + \frac{1}{4}(\beta_2 + \beta_3 + \beta_3 + \beta_4) + \frac{1}{4}(\gamma_1 + \gamma_2 + \gamma_4 + \gamma_3) + \bar{\varepsilon}_{\cdot\cdot 1}\end{aligned}$$

The estimated difference in the mean responses of treatments 3 and 1 is

$$\hat{\mu}_{\cdot\cdot 3} - \hat{\mu}_{\cdot\cdot 1} = \bar{y}_{\cdot\cdot 3} - \bar{y}_{\cdot\cdot 1} = \tau_3 - \tau_1 + \left[(\bar{\varepsilon}_{\cdot\cdot 3} - \bar{\varepsilon}_{\cdot\cdot 1}) + \frac{1}{4}(\beta_1 - \beta_2 - \beta_3 + \beta_4)\right]$$

Thus, the estimated difference between the mean responses from treatment III and treatment I would involve row effects. Thus, we have treatment effects confounded with row effects. This results from treatment I not appearing in row 1 but appearing twice in row 3 and from treatment III not appearing in row 2 but appearing twice in row 4. ■

Following the same reasoning, if a completely randomized design was used when a Latin square design was appropriate, the error of estimation would be inflated by both row and column effects that do not cancel out.

test for treatment effects

We can **test specific hypotheses concerning the parameters in our model.** In particular, we may wish to test the hypothesis of no difference among the t treatment means. This hypothesis can be stated in the form

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_1 = 0$$

The alternative hypothesis would be

$$H_a: \text{ At least one of the } \tau_k\text{s is not equal to zero.}$$

In terms of the treatment means, the hypotheses are

$$H_0: \mu_{..1} = \mu_{..2} = \cdots = \mu_{..t}$$

$$H_a: \text{ At least one } \mu_{..k} \text{ differs from rest.}$$

Our test statistic will be obtained by examining the model for a Latin square design and partitioning the total sum of squares to include terms for treatment effects, row effects, column effects, and random error effects.

total sum of squares

The **total sum of squares** of the measurements about their mean $\bar{y}_{...}$ is defined as before:

$$\text{TSS} = \sum_{ij} (y_{ijk} - \bar{y}_{...})^2$$

This sum of squares will be partitioned into four separate sources of variability: one due to the variability among treatments, one due to the variability among rows, one due to the variability among columns, and one due to the variability from all sources not accounted for by either treatment differences or block differences. We call this source of variability **error.** The **partition of TSS** follows.

error
partition of TSS

$$\text{TSS} = \sum_{i} \sum_{j} (y_{ijk} - \bar{y}_{...})^2$$

$$\text{TSS} = t\sum_{k} (\bar{y}_{..k} - \bar{y}_{...})^2 + t\sum_{i} (\bar{y}_{i..} - \bar{y}_{...})^2 + t\sum_{j} (\bar{y}_{.j.} - \bar{y}_{...})^2 + \text{SSE}$$

We will interpret the terms in the partition using the parameter estimates. The first quantity on the right-hand side of the equal sign measures the variability of the treatment means $\bar{y}_{..k}$ from the overall mean $\bar{y}_{...}$. Thus,

$$\text{SST} = t\sum_{k} (\bar{y}_{..k} - \bar{y}_{...})^2$$

between-treatment sum of squares

called the **between-treatment sum of squares,** is a measure of the variability in the y_{ijk}s due to differences in the treatment means. The second quantity,

$$\text{SSR} = t\sum_{i} (\bar{y}_{i..} - \bar{y}_{...})^2$$

between-rows sum of squares
between-columns of sum of squares

measures the variability between the row means $\bar{y}_{i..}$ and the overall mean. It is called the **between-rows sum of squares.** The third source of variability, referred to as the **between-columns sum of squares,** measures the variability between the column means $\bar{y}_{.j.}$ and the overall mean. It is given by

$$\text{SSC} = t\sum_{j} (\bar{y}_{.j.} - \bar{y}_{...})^2$$

TABLE 15.14
Analysis of variance table for a $t \times t$ Latin square design

Source	SS	df	MS	F
Treatments	SST	$t-1$	$\text{MST} = \text{SST}/(t-1)$	MST/MSE
Rows	SSR	$t-1$	$\text{MSR} = \text{SSR}/(t-1)$	MSR/MSE
Columns	SSC	$t-1$	$\text{MSC} = \text{SSC}/(t-1)$	MSC/MSE
Error	SSE	$(t-1)(t-2)$	$\text{MSE} = \text{SSE}/(t-1)(t-2)$	
Total	TSS	t^2-1		

The final source of variability, designated as the **sum of squares for error,** SSE, represents the variability in the y_{ijk}s not accounted for by the row, column, and treatment differences. It is given by

$$\text{SSE} = \text{TSS} - \text{SST} - \text{SSR} - \text{SSC} = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})$$

We can summarize our calculations in an AOV table, as given in Table 15.14.

The test statistic for testing

H_0: $\mu_{..1} = \mu_{..2} = \cdots = \mu_{..t}$ versus H_a: At least one $\mu_{..k}$ differs from the rest

or equivalently,

H_0: $\tau_1 = \tau_2 = \cdots = \tau_t = 0$ versus H_a: At least one τ_k differs from zero

is the ratio

$$F = \frac{\text{MST}}{\text{MSE}}$$

For our model,

$$E(\text{MSE}) = \sigma_\varepsilon^2 \quad \text{and} \quad E(\text{MST}) = \sigma_\varepsilon^2 + t\theta_T$$

where $\theta_T = 1/(t-1)\sum_k (\mu_{..k} - \mu_{...})^2$. When H_0 is true, $\mu_{..k} = \mu_{...}$ for all $k = 1, \ldots, t$, and, hence, $\theta_T = 0$. Thus, when H_0 is true, we would expect MST/MSE to be close to 1. However, under the research hypothesis, H_a, θ_T would be positive, since at least one of the differences $(\mu_{..k} - \mu_{...})$ is not 0. Thus, a large difference in the treatment means will result in a large value for θ_T. The expected value of MST will then be larger than the expected value of MSE, and we would expect $F = \text{MST}/\text{MSE}$ to be larger than 1. As a result, our test statistic F rejects H_0 when we observe a value of F larger than a value in the upper tail of the F distribution.

The above discussion leads to the following decision rule for a specified probability of a Type I error:

$$\text{Reject } H_0\text{: } \mu_{..1} = \mu_{..2} = \cdots = \mu_{..t} \text{ when } F = \text{MST}/\text{MSE} \text{ exceeds } F_{\alpha,\, \text{df}_1,\, \text{df}_2}$$

where $F_{\alpha,\, \text{df}_1,\, \text{df}_2}$ is from the F tables of Appendix Table 8 with α = specified value of the probability of a Type I error, $\text{df}_1 = \text{df}_{\text{MST}} = t - 1$, and $\text{df}_2 = \text{df}_{\text{MSE}} = (t-1)(t-2)$. Alternatively, we can compute the p-value for the observed value of the test statistic F_{obs} by computing

$$p\text{-value} = P(F_{\text{df}_1,\, \text{df}_2} > F_{\text{obs}}) = 1 - \boldsymbol{pf}(F_{\text{obs}}, t-1, (t-1)(t-2)).$$

where the F distribution with $\text{df}_1 = t - 1$ and $\text{df}_2 = (t-1)(t-2)$ is used to compute the probability. We would then compare the p-value to a selected value for the probability of a Type I error, with small p-values supporting the research hypothesis and large p-values failing to reject H_0.

EXAMPLE 15.5

The consumer-product rating organization decided to design the study of home air cleaners as a Latin square design using five homes and five months as blocking variables. The response variable is the CADR value obtained from a room air cleaner in a given home during a given month. Each brand of cleaner is observed in all five homes during all five months. The data from this study are given in Table 15.15. According to industry standards, a CADR value above 300 is considered excellent, and a CADR value below 100 is considered poor. Use these data to answer the following questions.

TABLE 15.15
CADR value for five brands of air cleaners in a 5×5 Latin square design

	Home						
Month	**1**	**2**	**3**	**4**	**5**	**Month Mean**	**Brand Mean**
M_1	$B_1(162)$	$B_2(89)$	$B_3(160)$	$B_4(146)$	$B_5(241)$	159.6	182.2
M_2	$B_2(115)$	$B_3(192)$	$B_4(164)$	$B_5(296)$	$B_1(142)$	181.8	139.8
M_3	$B_3(149)$	$B_4(273)$	$B_5(238)$	$B_1(227)$	$B_2(103)$	198.0	165.6
M_4	$B_4(229)$	$B_5(273)$	$B_1(175)$	$B_2(71)$	$B_3(119)$	173.4	229.0
M_5	$B_5(328)$	$B_1(205)$	$B_2(321)$	$B_3(208)$	$B_4(333)$	279.0	275.2
Home mean	196.6	206.4	211.6	189.6	187.6		

a. Write an appropriate statistical model for this experimental situation.
b. Conduct an analysis of variance to compare the mean CADRs for the five brands of air cleaners. Use $\alpha = .05$.

Solution **a.** The experiment was conducted as a Latin square design with $t = 5$ rows (months), $t = 5$ columns (homes), and $t = 5$ treatments (brands of air cleaners). An appropriate statistical model for this study is

$$y_{ijk} = \mu + \tau_k + \beta_i + \gamma_j + \varepsilon_{ijk} \text{ with } \quad i, j, k = 1, 2, 3, 4, 5$$

b. From the information in Table 15.15, the treatment means $\mu_{..k}$ are estimated by $\hat{\mu}_{..k} = \bar{y}_{..k}$, yielding

$$\hat{\mu}_{..1} = 182.2 \quad \hat{\mu}_{..2} = 139.8 \quad \hat{\mu}_{..3} = 165.6 \quad \hat{\mu}_{..4} = 229.0 \quad \hat{\mu}_{..5} = 275.2$$

From the above estimated treatment means, it appears that brand B_5 has a somewhat larger mean CADR value than brand B_4 and a considerably larger mean CADR value than the other three brands. From the data in Table 15.15, the sum of squares can be computed using the following formulas (note that $\bar{y}_{...} = 198.36$):

$$\begin{aligned}
\text{TSS} &= \sum_{ijk} (y_{ijk} - \bar{y}_{...})^2 \\
&= (162 - 198.36)^2 + (115 - 198.36)^2 + \cdots + (333 - 198.36)^2 \\
&= 139{,}372 \\
\text{SST} &= t \sum_{k} (\bar{y}_{..k} - \bar{y}_{...})^2 \\
&= 5[(182.2 - 198.36)^2 + (139.8 - 198.36)^2 + (165.6 - 198.36)^2 \\
&\quad + (229.0 - 198.36)^2 + (275.2 - 198.36)^2] = 58{,}034.16 \\
\text{SSR} &= t \sum_{i} (\bar{y}_{i..} - \bar{y}_{...})^2 \\
&= 5[(159.6 - 198.36)^2 + (181.8 - 198.36)^2 + (198.0 - 198.36)^2 \\
&\quad + (173.4 - 198.36)^2 + (279.0 - 198.36)^2] = 44{,}512.56
\end{aligned}$$

$$\text{SSC} = t\sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$$
$$= 5[(196.6 - 198.36)^2 + (206.4 - 198.36)^2 + (211.6 - 198.36)^2 + (189.6 - 198.36)^2 + (187.6 - 198.36)^2] = 2{,}177.76$$

By subtraction, we obtain the sum of squares error:

$$\text{SSE} = \text{TSS} - \text{SST} - \text{SSR} - \text{SSC}$$
$$= 139{,}372 - 58{,}034.16 - 44{,}512.56 - 2{,}177.76 = 34{,}647.52$$

The analysis of variance is summarized in Table 15.16.

TABLE 15.16
Analysis of variance for Example 15.5

Source	df	SS	MS	*F*	*p*-value
Month	4	44,512.56	11,128.14	3.85	.031
Home	4	2,177.76	544.44	.19	.940
Brand	4	58,034.16	14,508.54	5.02	.013
Error	12	34,647.52	2,887.29		
Total	24	139,372.00			

Note that the mean square for a source of variation in the AOV table is computed by dividing the sum of squares for that source by its degrees of freedom. The F test for differences in the five brands of air cleaners is $F = \text{MST/MSE}$. The computed value of $F = 5.02$ is greater than $F_{4,\,12,\,.05} = 3.26$, the tabulated F-value, based on $\text{df}_1 = 4$, $\text{df}_2 = 12$, and $\alpha = .05$. Therefore, we conclude that there is significant evidence (p-value = .013) of a difference in the mean CADR values for the five brands of air cleaners. It appears that brands B_4 and B_5 have higher mean CADR values than the other three brands. It is possible that brand B_5 has a higher mean CADR value than brand B_4. This could be confirmed by using a multiple-comparison procedure. ■

In order to validly make the inferences described in Example 15.5, it is necessary to verify that the conditions of independence, normality, and equal variances hold. This would involve a residual analysis using the residuals from the Latin square model—namely, $e_{ijk} = y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...}$. The condition of independence can be assessed only if there is a variable that allows us to sequence the residuals. Such variables are generally the order in which the measurements in the experiment were taken or a spatial relationship between the experimental units. If no such variable exists, the condition of independence needs to be confirmed subjectively by the researchers. The condition of normality can be assessed by a normal probability plot of the residuals and/or a test of normality of the residuals. The constancy of variance can be ascertained by plotting the residuals versus the predicted values of the observations, $\hat{y}_{ijk} = \bar{y}_{i..} + \bar{y}_{.j.} + \bar{y}_{..k} - 2\bar{y}_{...}$. If the spread in the residuals stays relatively constant with increasing $\hat{y}_{ijk}$, then the condition of constant variance would appear not to be violated. A residual analysis of the data from Example 15.5 is presented in Example 15.6.

EXAMPLE 15.6

The normal probability plot of the residuals and a plot of the residuals versus the predicted values are given here. Is there evidence that the conditions of normality and constant variance appear to be violated?

Solution From the normal probability plot, Figure 15.3(a), the plotted points are in close proximity to a straight line. The p-value for the test of normality is given to be p-value $>.10$, which indicates that there is not significant evidence of a violation of the normality condition. The plot of the residuals versus the fitted values, Figure 15.3(b), does not indicate a violation of the constant variance condition. Thus, the consumer-product testing organization can feel confident in publishing the conclusions from the AOV table.

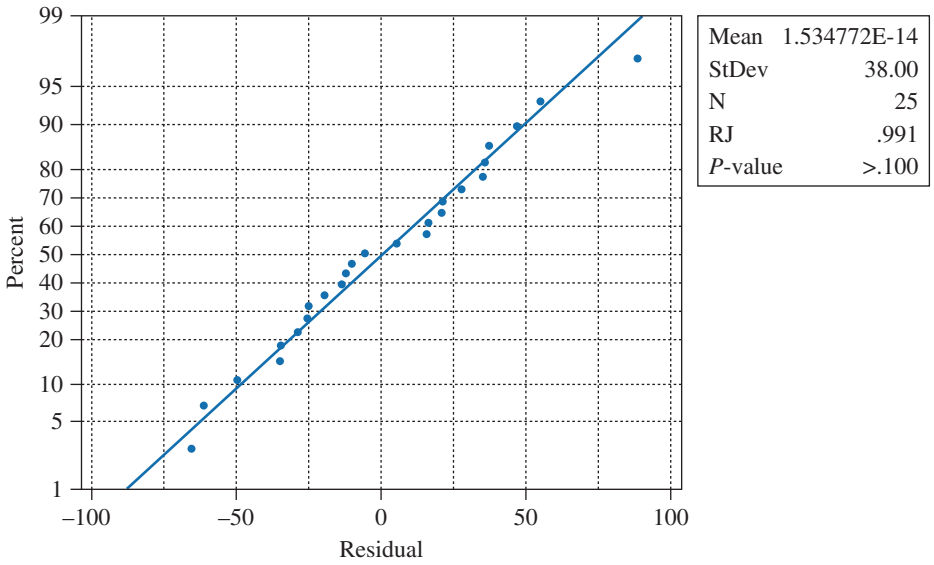


FIGURE 15.3(a)
Normal probability plot of air cleaner data

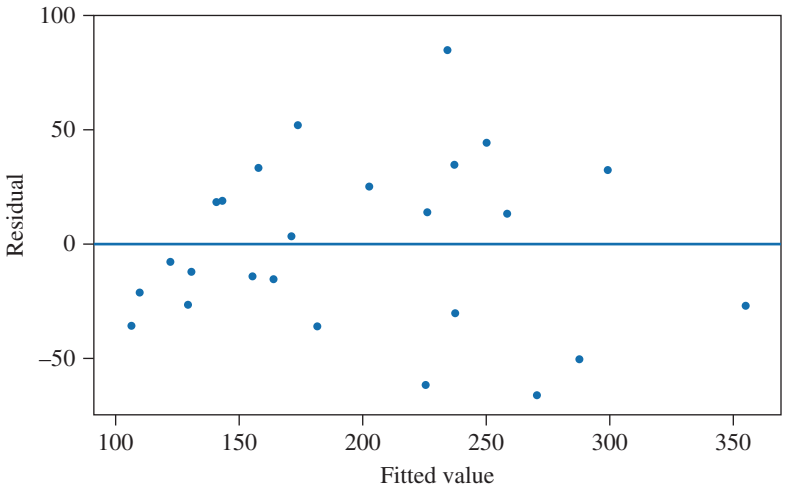


FIGURE 15.3(b)
Plot of residuals versus fitted values for air cleaner data

relative efficiency

RE(LS, CR)

The row and column effects are generally assessed only to determine whether or not accounting for the two extraneous sources of variability was efficient in reducing the variability in the experimental units. Thus, hypotheses about the row and column effects are not generally tested. As with the randomized block design, we can compare the efficiency of the Latin square design to that of the completely randomized design. We want to determine whether accounting for the row and column sources of variability has increased our precision for comparing treatment means in a given experiment. Let MSE_{LS} and MSE_{CR} denote the mean square errors for a Latin square design and a completely randomized design, respectively. The **relative efficiency** of the Latin square design compared to that of a completely randomized design is denoted **RE(LS, CR).** We can use the mean squares from the Latin square design—MSR, MSC, and MSE—to obtain the relative efficiency RE(LS, CR) by using the formula

$$\text{RE(LS, CR)} = \frac{\text{MSE}_{\text{CR}}}{\text{MSE}_{\text{LS}}} = \frac{\text{MSR} + \text{MSC} + (t-1)\text{MSE}}{(t+1)\text{MSE}}$$

When RE(LS, CR) is much larger than 1, we conclude that accounting for the row and/or column sources of variability was efficient, since many more observations would be required in a completely randomized design than in a Latin square design to obtain the same degree of precision in estimating the treatment means.

The following example will illustrate the calculations of the relative efficiency.

EXAMPLE 15.7

Refer to Example 15.5. Assess whether taking into account the two extraneous sources of variation, months and homes, was effective in increasing the precision of the analysis relative to a completely randomized design.

Solution From the AOV table in Example 15.5, we have MSR = MSMONTH = 11,128.14, MSC = MSHOME = 544.44, and MSE = 2,887.29. Thus, the relative efficiency of this Latin square design relative to a completely randomized design is given by

$$\begin{aligned}\text{RE(LS, CR)} &= \frac{\text{MSR} + \text{MSC} + (t-1)\text{MSE}}{(t+1)\text{MSE}} \\ &= \frac{11{,}128.14 + 544.44 + (5-1)(2{,}887.29)}{(5+1)(2{,}887.29)} = 1.34\end{aligned}$$

That is, approximately 34% more observations per treatment would be required in a completely randomized design to obtain the same precision in estimating the treatment means as with this Latin square design. The Latin square design has provided a considerable increase in the precision of estimation over a completely randomized design. However, this does not mean that both the row- and column-blocking factors are equally effective. In fact, it would appear from the relative sizes of the mean squares for months and for homes that the major portion of the gain in precision is from the month blocking factor. The differences in the means for the five homes are relatively small compared to the differences in the five monthly means. ■

EXAMPLE 15.8

To illustrate the output from a software package, the data from Example 15.5 were analyzed using the Minitab software. The output is given here.

```
General Linear Model: CADR versus MONTH, HOME, BRAND

Factor    Type     Levels  Values
MONTH     fixed         5  1, 2, 3, 4, 5
HOME      fixed         5  1, 2, 3, 4, 5
BRAND     fixed         5  1, 2, 3, 4, 5


Analysis of Variance for CADR, using Adjusted SS for Tests

Source   DF   Seq SS   Adj SS  Adj MS     F      P
MONTH     4    44513    44513   11128  3.85  0.031
HOME      4     2178     2178     544  0.19  0.940
BRAND     4    58034    58034   14509  5.02  0.013
Error    12    34647    34647    2887
Total    24   139372


S = 53.7334   R-Sq = 75.14%   R-Sq(adj) = 50.28%

Least Squares Means for CADR

BRAND  Mean    SE Mean
1      182.2    24.03
2      139.8    24.03
3      165.6    24.03
4      229.0    24.03
5      275.2    24.03

Unusual Observations for CADR

Obs    CADR      Fit  SE Fit  Residual  St Resid
 23 321.000 233.680  38.748    87.320      2.35 R

R denotes an observation with a large standardized residual.
```

Note that in the output from Minitab, there are two types of sums of squares listed: Seq SS and Adj SS. In nearly all situations, the Adj SS will be the sum of squares that will be used in assessing treatment differences. Also, observation 23—month = 5, home = 3, brand = B_2—has been identified as an unusual observation. This data point can be seen in the two plots of the residuals displayed in Figure 15.3. Although this observation has a moderately large standardized residual, 2.35, it is not large enough to cause too much concern about its impact on the validity of the F test. ■

15.4 Factorial Treatment Structure in a Randomized Complete Block Design

In Chapter 14, we discussed a completely randomized design with a factorial treatment structure in which the response y is observed at all factor–level combinations of the independent variables. The factor–level combinations of the independent variables (treatments) were randomly assigned to the experimental units in order to investigate the effects of the factors on the response.

Sometimes the objectives of a study are such that we wish to investigate the effects of certain factors on a response while blocking out certain other extraneous sources of variability. Such situations require a block design with treatments from a factorial treatment structure. We will draw on our knowledge of block designs (randomized block designs and Latin square designs) to effectively block out the extraneous sources of variability in order to focus on the effects of the factors on the response of interest. This can be illustrated with the following example.

EXAMPLE 15.9

A nutritionist wants to study the percentage of protein content in bread made from three new types of flours and baked at three different temperatures. She would like to bake 3 loaves of bread from each of the nine flour–temperature combinations for a total of 27 loaves from which the percentage of protein would be determined. However, she is able to bake only 9 loaves on any given day. Propose an appropriate experimental design.

Solution Because nine loaves can be baked on a given day, it would be possible to run a complete replication of the 3×3 factorial treatment structure on three different days to obtain the desired number of observations. The design is shown in Table 15.17.

TABLE 15.17
Protein content data

	Day								
	1			2			3		
Flour	Temperature			Temperature			Temperature		
Type	1	2	3	1	2	3	1	2	3
A	5.8	4.6	4.6	11.4	5.2	5.2	10.5	9.7	4.7
B	8.4	5.4	4.7	7.5	7.9	7.2	14.6	7.9	6.9
C	16.0	5.2	4.2	17.8	7.0	6.3	16.9	11.5	7.2

Note that this design is really a randomized block design, where the blocks are days and the treatments are the nine factor–level combinations of the 3×3 factorial treatment structure. So, with the randomized block design, we are able to block or filter out the variability due to the nuisance variable, days, while comparing the treatments. Because the treatments are factor–level combinations from a factorial treatment structure, we can examine the effects of the two factors (flour and temperature) on the response while filtering out the day-to-day variability.

The analysis of variance for this design follows from our discussions in Sections 14.3 and 15.2. ■

The model for a randomized complete block design with an $a \times b$ factorial treatment structure is given here:

$$y_{ijk} = \mu + \beta_i + \tau_j + \gamma_k + \tau\gamma_{jk} + \varepsilon_{ijk}$$

where the terms in the model are defined as follows:

- y_{ijk}: Response from the experimental unit in the ith block receiving the jth level of factor A and kth level of factor B.
- μ: Overall mean, an unknown constant.
- β_i: Effect due to the ith block, an unknown constant.
- τ_j: Effect due to the jth level of factor A, an unknown constant.
- γ_k: Effect due to the kth level of factor B, an unknown constant.
- $\tau\gamma_{jk}$: Interaction effect of the jth level of factor A with the kth level of factor B, an unknown constant.
- ε_{ijk}: Random error associated with the response from the experimental unit in the ith block receiving the jth level of factor A and the kth level of factor B. We require that the ε_{ijk}s have a normal distribution with a mean of 0 and a common variance of σ_ε^2. In addition, the ε_{ijk}s must be independently distributed.

TABLE 15.18
AOV for a randomized complete block design with two factors

Source	df	SS	MS	*F*
Blocks	$r - 1$	SSBL	MSBL	MSBL/MSE
Treatments	$ab - 1$	SST	MST	MST/MSE
A	$a - 1$	SSA	MSA	MSA/MSE
B	$b - 1$	SSB	MSB	MSB/MSE
AB	$(a - 1)(b - 1)$	SSAB	MSAB	MSA/MSE
Error	$(ab - 1)(r - 1)$	SSE	MSE	
Total	$abr - 1$	SST		

The conditions given above for our model can be shown to imply that the responses, y_{ijk}, have a normal distribution with mean

$$\mu_{ijk} = E(y_{ijk}) = \mu + \beta_i + \tau_j + \gamma_k + \tau\gamma_{jk}$$

and variance σ_ε^2.

The sums of squares can be computed using the following formulas:

$$\text{TSS} = \sum_{ijk} (y_{ijk} - \bar{y}_{...})^2$$

$$\text{SST} = r\sum_{jk} (\bar{y}_{.jk} - \bar{y}_{...})^2$$

$$\text{SSBL} = ab\sum_{i} (\bar{y}_{i..} - \bar{y}_{...})^2$$

$$\text{SSA} = rb\sum_{j} (\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$\text{SSB} = ra\sum_{k} (\bar{y}_{..k} - \bar{y}_{...})^2$$

$$\text{SSAB} = r\sum_{jk} (\bar{y}_{.jk} - \bar{y}_{...})^2 - \text{SSA} - \text{SSB}$$

By subtraction, we obtain the sum of squares error:

$$\text{SSE} = \text{TSS} - \text{SSBL} - \text{SSA} - \text{SSB} - \text{SSAB}$$

Furthermore, we have

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB}$$

The AOV table for a randomized complete block design with r blocks and two factors, factor A with a levels and factor B with b levels, is given in Table 15.18.

EXAMPLE 15.10

Construct an analysis of variance table for the experiment described in Example 15.9.

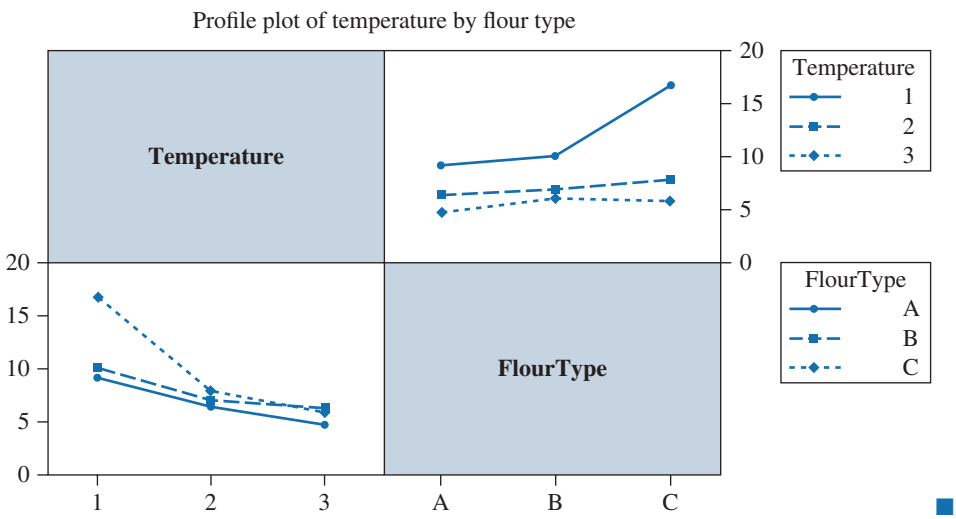
Solution The following output from Minitab is given here.

```
General Linear Model: Protein% versus Day, Temperature, FlourType

Factor       Type  Levels  Values
Day          fixed      3  1, 2, 3
Temperature  fixed      3  1, 2, 3
FlourType    fixed      3  A, B, C


Analysis of Variance for Protein%, using Adjusted SS for Tests

Source                 DF   Seq SS   Adj SS   Adj MS      F      P
Day                     2   53.479   53.479   26.739   9.39  0.002
Temperature             2  204.156  204.156  102.078  35.85  0.000
FlourType               2   54.376   54.376   27.188   9.55  0.002
Temperature*FlourType   4   56.913   56.913   14.228   5.00  0.008
Error                  16   45.555   45.555    2.847
Total                  26  414.479
```

```
Least Squares Means for Protein%

Temperature             Mean    SE Mean
FlourType
Temp.     FlourType
1              A       9.233     0.9742
1              B      10.167     0.9742
1              C      16.900     0.9742
2              A       6.500     0.9742
2              B       7.067     0.9742
2              C       7.900     0.9742
3              A       4.833     0.9742
3              B       6.267     0.9742
3              C       5.900     0.9742
-------------------------------------
1                     12.100     0.5625
2                      7.156     0.5625
3                      5.667     0.5625
               A       6.856     0.5625
               B       7.833     0.5625
               C      10.233     0.5625
```

From the output, we can observe that there is a significant interaction (p-value = .008) between temperature and flour type. This interaction is displayed in the profile plot in Figure 15.4.

FIGURE 15.4
Profile plot displaying the interaction between temperature and flour type



Because of the significant interaction in Example 15.10, we would compare the mean percentages of protein for the three flour types separately at each temperature. Alternatively, we could compare the mean percentages of protein for the three temperatures separately at each level of flour type.

The Tukey W procedure could be used to obtain simultaneous confidence intervals on the differences in the mean responses for pairs of flour types at a fixed temperature $(\mu_{.jk} - \mu_{.j'k})$. These confidence intervals are given by

$$\bar{y}_{.jk} - y_{.j'k} \pm W \quad \text{where} \quad W = q_\alpha(t, \nu)\sqrt{\frac{s_w^2}{r}}$$

with $s_w^2 = \text{MSE}$, $t = ab$, $\nu = \text{df}_{\text{error}}$, and $q_\alpha(t, \nu)$ is the value from Table 10 in the Appendix.

Also, any pair of treatment means with $|\bar{y}_{.jk} - \bar{y}_{.j'k'}| \geq W$ would imply that there is significant evidence that the treatment means, $\mu_{.jk} - \mu_{.j'k'}$, are different.

EXAMPLE 15.11

For the experiment in Example 15.9, determine which pairs of treatments have significantly different means.

Solution Because there was significant evidence of an interaction, the comparisons of the mean responses for flour types are made separately at each temperature. After determining that $r = 3, a = b = 3, t = 9, \nu = 16$, and MSE = 2.847, for $\alpha = .05$ Table 10 in the Appendix yields $q\alpha(t, \nu) = q_{.05}(9, 16) = 5.03$. Thus, we have

$$W = q_{\alpha}(t, \nu)\sqrt{\frac{s_w^2}{r}} = (5.03)\sqrt{\frac{2.847}{3}} = 4.9$$

As a result, any pair of treatment means having a difference between corresponding sample means exceeding 4.9 would be declared significantly different. The pairwise differences are displayed in Table 15.19.

TABLE 15.19
Pairwise comparisons of flour types for each temperature level

Temperature	Flour Type Difference	$\lvert\bar{y}_{\cdot jk} - \bar{y}_{\cdot j'k'}\rvert$	Conclusion
1	A versus B	.934	Not significantly different
1	A versus C	7.667	Significantly different
1	B versus C	6.733	Significantly different
2	A versus B	.567	Not significantly different
2	A versus C	1.4	Not significantly different
2	B versus C	.833	Not significantly different
3	A versus B	1.434	Not significantly different
3	A versus C	1.067	Not significantly different
3	B versus C	.367	Not significantly different

■

15.5 A Nonparametric Alternative—Friedman's Test

In a randomized block experiment with b blocks and t treatments, when the condition that the residuals have a normal distribution is violated, one alternative is to attempt a transformation of the data. In some situations, it is not possible to determine an appropriate transformation. In a more extreme situation, the response variables may not have a continuous scale but only be ordinal. That is, the experimental units are simply ordered without a scale. This type of response often occurs when the responses are obtained as ratings by experts, such as in food tasting or sports in which judges are used to assess the performance of the athletes. In both the case of nonnormally distributed data and the case of purely ordinal data, an appropriate test of no treatment difference is the Friedman test. The conditions under which the Friedman test is valid are listed here.

1. The experimental design is a randomized block design, with the t treatments randomly assigned to exactly one experimental unit per block, yielding $N = tb$ responses.
2. The N responses, y_{ij}, are mutually independent.
3. The N responses are related by the model $y_{ij} = \theta + \tau_i + \beta_j + \varepsilon_{ij}$, where θ is the overall median, τ_i is an effect due to the ith treatment, β_j is an effect due to the jth block, and the N ε_{ij}s are a random sample from a continuous distribution with a median equal to 0.

Note that if we further required that the ε_{ij}s have a normal distribution, then we would have the same requirements as in the standard AOV model.

The hypotheses being tested by Friedman's test involve the medians of the population distributions, whereas in the standard AOV we are testing hypotheses concerning the population means. In the normal distribution, the mean and median are the same and hence the equivalence between the two sets of hypotheses. The Friedman test is requires that the distributions of the responses differ only with respect to their medians and that all other aspects of the distributions be the same. This is equivalent to the distributional requirements in the standard AOV hypotheses, where we required the distributions of the residuals to reside within a normal family of distributions and to have the same variances. Thus, the distributions could differ only with respect to their medians.

In Chapter 8, we introduced the Kruskal–Wallis test for comparing t treatments when the experimental design was completely randomized. The Friedman test very similar to the Kruskal–Wallis test in that the procedures for the Friedman test replace the observed responses, y_{ij}, with their ranks. The difference between the two procedures lies in how the data values are ranked. The Kruskal–Wallis test ranks the N data values as a whole, thus replacing the responses, y_{ij}, with the integers $1, 2, \ldots, N$. The Friedman test obtains a separate ranking of the data values within each of the b blocks. Thus, the data values in each block are replaced with the integers $1, 2, \ldots, t$.

The steps for conducting the Friedman test are as follows:

1. Order the t observations from smallest to largest separately within each of the b blocks.
2. Replace the observations with R_{ij}, the ranks of y_{ij} in the joint ranking of the data values $y_{1j}, y_{2j}, \ldots, y_{tj}$ in the jth block.
3. Compute the sum of the ranks and then the mean rank for the ith treatment:

$$R_{i.} = \sum_{j=1}^{b} R_{ij} \quad \text{and} \quad \overline{R}_{i.} = \frac{R_{i.}}{b}$$

   Thus, $R_{1.}$ is the sum of the ranks of the b observations on treatment 1, and $\overline{R}_{1.}$ is the average rank of the observations on treatment 1.

4. The Friedman test is then given by

$$\text{FR} = \frac{12b}{t(t+1)} \sum_{i=1}^{t} \left(\overline{R}_{i.} - \frac{t+1}{2} \right)^2 = \left(\frac{12}{bt(t+1)} \sum_{i=1}^{t} R_{i.}^2 \right) - 3b(t+1)$$

   where $\dfrac{t+1}{2}$ is the average rank within each of the b blocks.

To test the research hypothesis that the t treatments do not have the same median—that is, to test H_0: $\tau_1 = \tau_2 = \cdots = \tau_t$ versus H_a: $\tau_1, \tau_2, \ldots, \tau_t$ are not all equal.—

$$\text{Reject } H_0 \text{ if } \text{FR} \geq \text{FR}_\alpha$$

where the critical value FR_α is selected to achieve a type I error rate of α. Values of FR_α can be obtained from the book *Nonparametric Statistical Methods* (Hollander and Wolfe, 1999). An approximation based on large sample theory is to

$$\text{Reject } H_0 \text{ if } \text{FR} \geq \chi^2_{\alpha, t-1}$$

where $\chi^2_{\alpha, t-1}$ is the upper α percentile from the chi-square distribution, Table 7 in the Appendix.

EXAMPLE 15.12

The paper ***"Physiological Effects During Hypnotically Requested Emotions"* (Damaser, Shor, and Orue, 1963)** reported the following data on skin potential (millivolts) when the emotions of fear, happiness, depression, and calmness were reported from each of eight subjects. In this study, the subjects serve as the blocks and the treatments are the four emotions. Perform a preliminary analysis of the data (shown in Table 15.20) to determine if the normal-based procedures can be applied. Then apply the Friedman test at the $\alpha = .05$ level to determine if there is a difference in the median skin potentials of the four emotions. Finally, compare the results from the two methods of testing for skin potential differences across the four emotions.

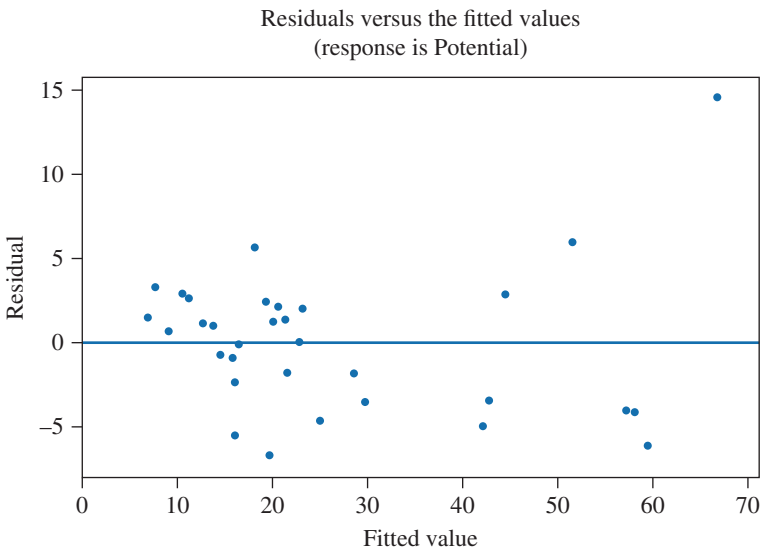
TABLE 15.20
Skin potential readings by emotion

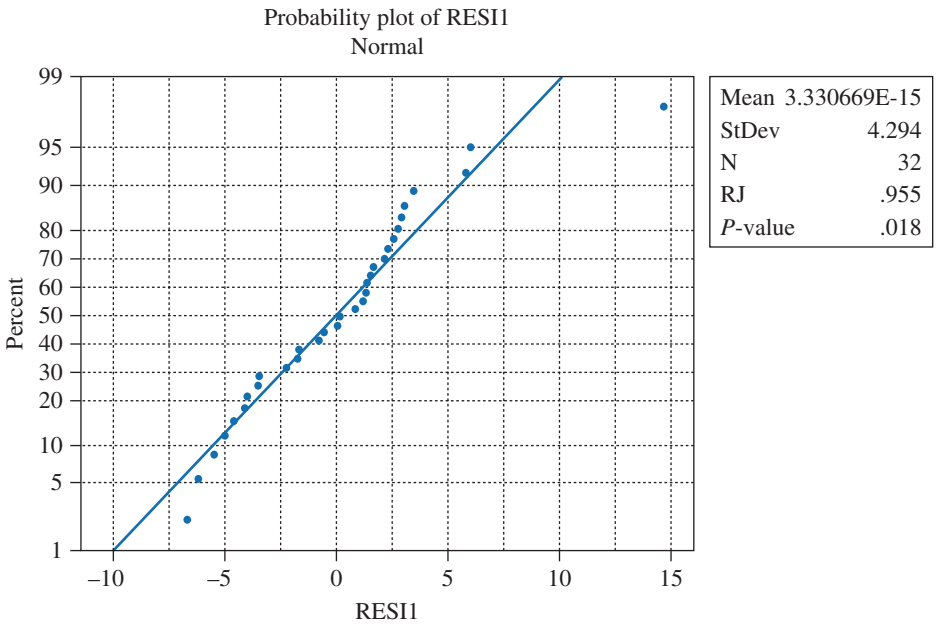
	Subjects (Blocks)							
Emotion	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Fear	26.1	81.0	10.5	26.6	12.9	57.2	25.0	20.3
Happiness	22.7	53.2	9.7	19.6	13.8	47.1	13.6	23.6
Depression	22.5	53.7	10.8	21.1	13.7	39.2	13.7	16.3
Calmness	22.6	53.1	8.3	21.6	13.3	37.0	14.8	14.8

Solution The responses were analyzed using the normal-based procedures, yielding the following Minitab output.

```
Two-way ANOVA: Potential versus Blocks, Emotion

Source    DF       SS       MS        F        P
Blocks     7  8465.80  1209.40    44.43    0.000
Emotion    3   433.28   144.43     5.31    0.007
Error     21   571.61    27.22
Total     31  9470.69

S = 5.217  R-Sq = 93.96%  R-Sq(adj) = 91.09%
```

The following residual plots were obtained also.





The F test from the AOV table yields a p-value = .007 for testing the difference in the mean skin potentials among the four emotions. This would indicate that there is significant evidence of a difference in the mean skin potentials among the four emotions. However, an examination of the residuals should be made prior to placing much confidence in this conclusion. From the normal probability plot of the residuals, it would appear there is a violation of the requirement that the residuals have a normal distribution. The test of normality has a p-value = .018, which confirms our observation from the plot. The next step is to test for a difference in median skin potentials using the Friedman test.

The skin potential readings were ranked from smallest to largest separately for each subject, with the smallest value receiving a ranking of 1 and the largest value receiving a ranking of $t = 4$. The rankings are given in Table 15.21.

From the rankings in Table 15.21, the Friedman test result is calculated as follows:

$$\text{FR} = \frac{12(b)}{t(t+1)}\sum_{i=1}^{t}\left(\overline{R}_{i.} - \frac{t+1}{2}\right)^2$$

$$\text{FR} = \frac{12(8)}{4(4+1)}\sum_{i=1}^{4}\left(\overline{R}_{i.} - \frac{4+1}{2}\right)^2$$

$$\text{FR} = 4.80\sum_{i=1}^{4}(\overline{R}_{i.} - 2.5)^2$$

$$= 4.8[(3.375 - 2.5)^2 + (2.5 - 2.5)^2 + (2.375 - 2.5)^2 + (1.75 - 2.5)^2]$$

$$= 6.45$$

Reject H_0 if $\text{FR} \geq \chi^2_{t-1,\alpha} = \chi^2_{3,.05} = 7.815$

FR = 6.45 < 7.815 and p-value = $Pr[\chi^2_3 \geq 6.45]$ = .092. Therefore, we fail to reject H_0 and conclude there is not significant evidence of a difference in the median skin potentials for the four emotions. This conclusion differs from the conclusion

TABLE 15.21
Ranks of treatments within each subject

	Subjects (Blocks)								Sum of Ranks	Mean Rank
Emotion	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	$R_{i.}$	R_i
Fear	4	4	3	4	1	4	4	3	27	3.375
Happiness	3	2	2	1	4	3	1	4	20	2.5
Depression	1	3	4	2	3	2	2	2	19	2.375
Calmness	2	1	1	3	2	1	3	1	14	1.75

reached using the AOV F test, where a significant difference was found in the mean skin potentials across the four emotions. An examination of the residuals reveals a few extreme values, which may have caused of the difference in the two conclusions. The skin potentials for fear for subject 2 and subject 6 were much larger than the values obtained from the other six subjects. These two large skin potentials would result in an inflated value for the mean skin potential for the fear emotion. The influence of these two values is greatly moderated in the Friedman test and hence the difference in the two conclusions. ■

15.6 RESEARCH STUDY: Control of Leatherjackets

Adult leatherjackets damage lawns by feeding on grass roots. A description of the types of problems resulting from these insects was given in Section 15.1. A study was designed to evaluate several proposed treatments for reducing the impact of leatherjackets on lawns.

Collecting the Data

The following experiment is described in the book *A Handbook of Small Data Sets* (Hand et al., 1993). It involved a control and four potential chemicals to eliminate the leatherjackets. Initially, the researchers were planning on evaluating the four new treatments on lawns at their research center. However, in order to broaden the level of inference of their study, they wanted to evaluate the chemicals on a variety of soils and terrains. Thus, plots of land at six different sites were selected for use in the experiment. A convenient way to conduct the experiment would be to use the same chemical at all test sites at a given location. However, this would result in the confounding of the effectiveness of the chemical with the location of the test sites. Therefore, the following experimental protocol was implemented. Within each of the six plots, there were 12 test sites, with 2 test sites randomly assigned to each of four treatments and 4 test sites randomly assigned to the control. A week after applying the treatments to the test sites, the researchers returned to the test sites and counted the number of surviving leatherjackets on each of the 72 test sites. The researchers were interested in determining if the average numbers of leatherjackets on the test sites receiving the four treatments were less than the average numbers on the control sites. Furthermore, they wanted to determine if there were differences in the four treatments relative to their average counts. The data collected during the experiment were given in Table 15.1. The treatment and block means are presented in Table 15.22.

TABLE 15.22
Mean leatherjacket counts on test sites

		Treatment				
Plot	**Control**	**1**	**2**	**3**	**4**	**Block Mean**
1	39.5	9.5	14.5	8	12.5	20.58
2	26.5	17.5	23.0	5.5	2.5	16.92
3	31.75	8.5	11.0	8.0	4.5	15.92
4	44.8	21.5	6.5	5.0	1.0	20.58
5	28.25	12.5	12.0	4.0	3.5	14.75
6	48.5	26.0	10.0	14.0	5.5	25.42
Treatment mean	36.54	15.92	12.83	7.42	4.92	19.03

Analyzing the Data

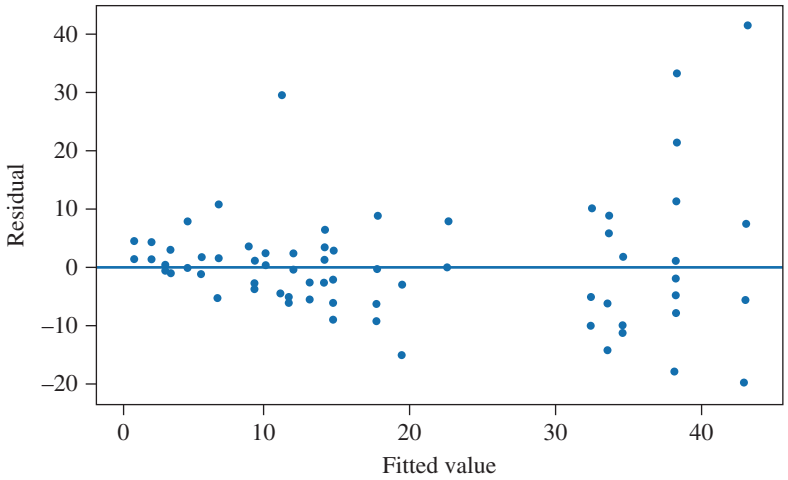
This is a randomized block experiment with $t = 5$ treatments. The blocks are the six plots of land, and the treatments are the four chemical pesticides and one control. Referring to Table 15.1, the model for this experiment would be

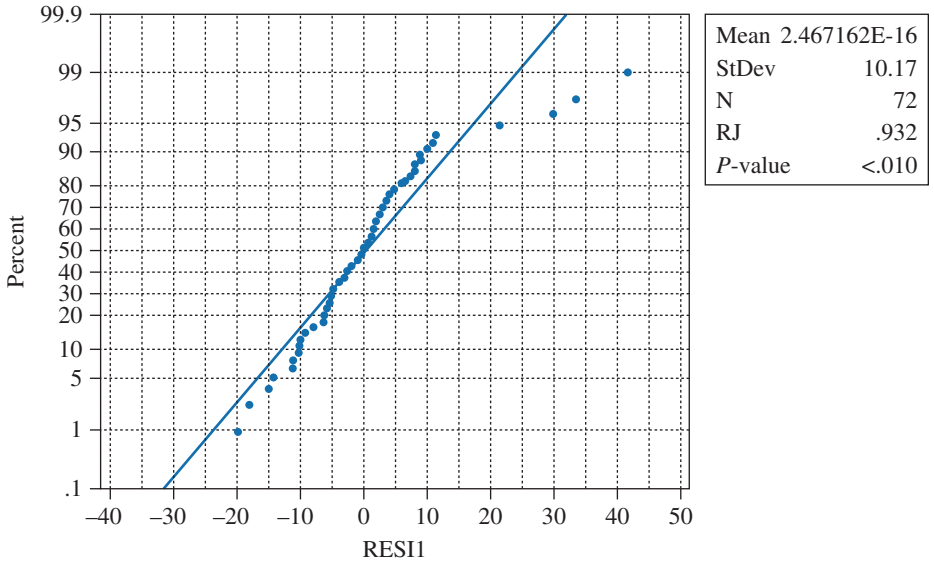
$$C_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}$$

with $i = 1, 2, 3, 4, 5$; $j = 1, 2, 3, 4, 5, 6$; $k = 1, 2, 3, 4$ for $i = 1$;
and $k = 1, 2$ for $i = 2, 3, 4, 5$

where C_{ijk} is the leatherjacket count on the kth test site in block j receiving treatment i. The data were analyzed using the above model, yielding the following AOV table and residual plots.

```
General Linear Model: Count versus Block, Treatment

Factor       Type     Levels    Values
Block        fixed         6    1, 2, 3, 4, 5, 6
Treatment    fixed         5    CNT, TRT1, TRT2, TRT3, TRT4

Analysis of Variance for Count, using Adjusted SS for Tests

Source       DF     Seq SS    Adj SS   Adj MS       F       P
Block         5      937.1     937.1    187.4    1.58   0.179
Treatment     4    11945.6   11945.6   2986.4   25.19   0.000
Error        62     7349.3    7349.3    118.5
Total        71    20231.9

S = 10.8874    R-Sq = 63.67%    R-Sq(adj) = 58.40%

Unusual Observations for Count

Obs      Count        Fit    SE Fit    Residual    St Resid
  2    59.0000    38.0972    3.6291     20.9028        2.04 R
 20    40.0000    10.7222    4.2556     29.2778        2.92 R
 37    71.0000    38.0972    3.6291     32.9028        3.21 R
 61    84.0000    42.9306    3.6291     41.0694        4.00 R

R denotes an observation with a large standardized residual.
```

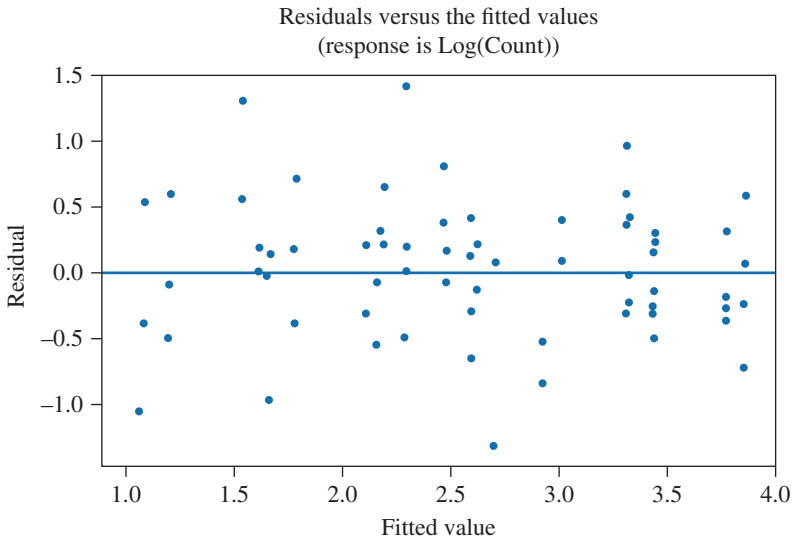


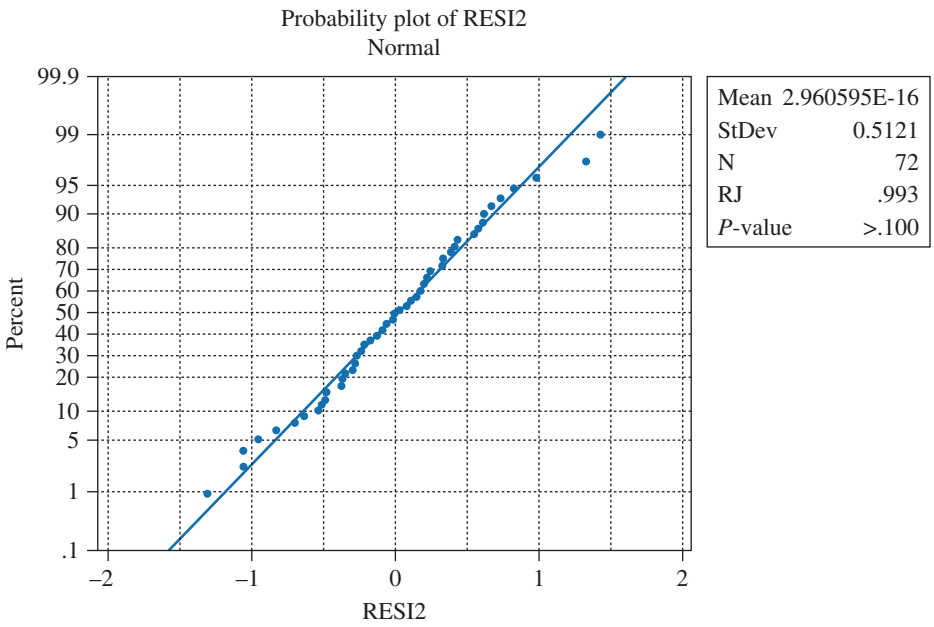


From the plot of the residuals versus the fitted values, it would appear that the variances are increasing with increasing fitted values. Also, the normal probability plot and the p-value $< .01$ for the test of normality both indicate that the conditions for using the F test in the AOV table are not satisfied. The output from Minitab also indicates that four test sites have large standardized residuals.

Thus, the conditions for validly using the F test to evaluate the differences in the mean counts for the five treatments do not appear to hold. Because the response variable is a count of the number of leatherjackets, two transformations are strongly suggested, the square root and log transformations. Both of these transformations were applied to the data, and the log transformation was the more effective in producing residuals having a normal distribution with constant variance. The following AOV table and residual plots were thus obtained.

```
General Linear Model: Log(Count) versus Block, Treatment

Factor     Type     Levels  Values
Block      fixed         6  1, 2, 3, 4, 5, 6
Treatment  fixed         5  CNT, TRT1, TRT2, TRT3, TRT4


Analysis of Variance for Log(Count), using Adjusted SS for Tests

Source     DF   Seq SS   Adj SS   Adj MS      F      P
Block       5   3.2501   3.2501   0.6500   2.16  0.069
Treatment   4  48.2335  48.2335  12.0584  40.15  0.000
Error      62  18.6209  18.6209   0.3003
Total      71  70.1045


S = 0.548030   R-Sq = 73.44%   R-Sq(adj) = 69.58%

  Unusual Observations for Log(Count)

  Obs  Log(Count)      Fit   SE Fit  Residual  St Resid
   11     2.83321  1.52253  0.21421   1.31068      2.60 R
   20     3.68888  2.27961  0.21421   1.40927      2.79 R
   47     0.00000  1.06017  0.21421  -1.06017     -2.10 R
   48     0.00000  1.06017  0.21421  -1.06017     -2.10 R
   68     1.38629  2.69892  0.21421  -1.31263     -2.60 R

  R denotes an observation with a large standardized residual.
```





From the preceding plots and with a p-value $> .10$, the conditions of normality and constant variance appear to hold for the transformed response, Log(Count). An examination of the AOV table reveals a p-value $< .0001$; thus, there is significant evidence of a difference in the five treatments. To further explore this difference, Tukey's W was applied to the treatment means with the following results.

```
Tukey Simultaneous Tests
Response Variable Log(Count)
All Pairwise Comparisons among Levels of Treatment
Treatment = CNT  subtracted from:

             Difference        SE of             Adjusted
Treatment     of Means   Difference   T-Value     P-Value
TRT1            -0.844       0.1938     -4.35      0.0005
TRT2            -1.148       0.1938     -5.92      0.0000
TRT3            -1.662       0.1938     -8.58      0.0000
TRT4            -2.240       0.1938    -11.56      0.0000

Treatment = TRT1  subtracted from:

             Difference        SE of             Adjusted
Treatment     of Means   Difference   T-Value     P-Value
TRT2            -0.304       0.2237    -1.358      0.6563
TRT3            -0.818       0.2237    -3.657      0.0047
TRT4            -1.396       0.2237    -6.239      0.0000

Treatment = TRT2  subtracted from:

             Difference        SE of             Adjusted
Treatment     of Means   Difference   T-Value     P-Value
TRT3            -0.514       0.2237    -2.299      0.1592
TRT4            -1.092       0.2237    -4.881      0.0001

Treatment = TRT3  subtracted from:

             Difference        SE of             Adjusted
Treatment     of Means   Difference   T-Value     P-Value
TRT4           -0.5778       0.2237    -2.583      0.0861
```

TABLE 15.23
Comparison of mean leatherjacket counts by treatment

	Treatment			
Control	**1**	**2**	**3**	**4**
36.54	15.92	12.83	7.42	4.92
A	B	B		
		C	C	
			D	D

The preceding output provides a pairwise comparison of the four new chemical treatments and a comparison of the treatments versus the control test sites. Using an experimentwise Type I error rate of $\alpha = .05$, the above p-values reveal that the mean for the control was significantly greater than all four treatment means. Next, examining the four treatment means, the mean for treatment 1 is not significantly different from that for treatment 2 but is significantly different from those for treatments 3 and 4. Treatment 2 is not significantly different from treatment 3 but is significantly different from treatment 4. Finally, treatment 3 is not significantly different from treatment 4. We can summarize these results as shown in Table 15.23.

The researchers plan on examining other potential chemicals for controlling insect infestations in residential lawns. A question of interest is whether it would be necessary to use all six locations in future experiments or if a single location would suffice. Using the data from the current study, the relative efficiency of using the six locations as the levels of a blocking factor compared to just running the experiment as a completely randomized design is computed as follows:

$$\begin{aligned} \text{RE(RCB, CR)} &= \frac{(b-1)\text{MSB} + b(t-1)\text{MSE}}{(bt-1)\text{MSE}} \\ &= \frac{(6-1)(187.4) + 6(5-1)(118.5)}{((6)(5)-1)(118.5)} = 1.10 \end{aligned}$$

Thus, it would take 10% more observations in a completely randomized design to achieve the same level of precision in estimating the treatment means as was achieved in the randomized complete block design:

15.7 Summary and Key Formulas

In this chapter, we discussed the analysis of variance presented for several different experimental designs and treatment structures. The designs considered were the randomized complete block design and the Latin square design. These designs illustrated how we can minimize the effect of undesirable variability from extraneous variables so as to obtain more precise comparisons among treatment means. The factorial treatment structure is useful in investigating the effect of one or more factors on an experimental response. Factorial treatments can be used in completely randomized, randomized complete block, and Latin square designs. Thus, an experimenter may wish to examine the effects of two or more factors on a response while blocking out one or more extraneous sources of variability.

For each design discussed in this chapter, we presented a description of the design layout (including arrangement of treatments), potential advantages and disadvantages, a model, and the analysis of variance. Finally, we discussed how one could conduct multiple comparisons between treatment means for each of these designs.

We discussed the importance of examining whether the conditions of independence, normality, and equal variance were satisfied in a given experimental setting. In the randomized complete block design, an alternative to the AOV F test, the Friedman text, should be implemented when the condition of normality is violated; otherwise, the level of the AOV F test may be incorrect.

Key Formulas

1. One factor in a randomized complete block design

   Model: $y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}; i = 1, \ldots, t; j = 1, \ldots, b$

   Sum of Squares:

   Total $\quad \text{TSS} = \sum_{ij}(y_{ij} - \bar{y}_{..})^2$

   Treatment $\quad \text{SST} = b\sum_i(\bar{y}_{i.} - \bar{y}_{..})^2$

   Block $\quad \text{SSB} = t\sum_j(\bar{y}_{.j} - \bar{y}_{..})^2$

   Error $\quad \text{SSE} = \sum_{ij}(e_{ij})^2 = \sum_{ij}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 = \text{TSS} - \text{SST} - \text{SSB}$

2. Relative efficiency of a randomized complete block design

$$\text{RE(RCB, CR)} = \frac{(b-1)\text{MSB} + b(t-1)\text{MSE}}{(bt-1)\text{MSE}}$$

3. One factor in a Latin square design

   Model: $y_{ijk} = \mu + \tau_k + \beta_i + \gamma_j + \varepsilon_{ijk}; i = j = k = 1, \ldots, t$

   Sum of Squares:

   Total $\quad \text{TSS} = \sum_{ijk}(y_{ijk} - \bar{y}_{...})^2$

   Treatment $\quad \text{SST} = t\sum_k(\bar{y}_{..k} - \bar{y}_{...})^2$

   Row $\quad \text{SSR} = t\sum_i(\bar{y}_{i..} - \bar{y}_{...})^2$

   Column $\quad \text{SSC} = t\sum_j(\bar{y}_{.j.} - \bar{y}_{...})^2$

   Error $\quad \text{SSE} = \text{TSS} - \text{SST} - \text{SSR} - \text{SSC}$

4. Relative efficiency of a Latin square design

$$\text{RE(LS, CR)} = \frac{\text{MSR} + \text{MSC} + (t-1)\text{MSE}}{(t+1)\text{MSE}}$$

5. Friedman's test in a randomized complete block design

$$\text{FR} = \frac{12b}{t(t+1)}\sum_{i=1}^{t}\left(\bar{R}_{i.} - \frac{t+1}{2}\right)^2$$

15.8 Exercises

15.2 Randomized Complete Block Design

Ag. **15.1** A horticulturist is designing a study to investigate the effectiveness of five methods for the irrigation of blueberry shrubs. The methods are surface, trickle, center pivot, lateral move, and subirrigation. There are 10 blueberry farms available for the study, representing a wide variety of types of soil, terrains, and wind gradients. The horticulturist wants to use each of the five methods of irrigation on all 10 farms to moderate the effect of the many extraneous sources of variation that may impact the blueberry yields. On each farm, five 1-acre plots are randomly selected, and a method of irrigation is randomly assigned to each plot. The response variable will be the weight of the harvested fruit from each of the plots of blueberry shrubs.

- **a.** Show the details of how you would randomly assign the five methods of irrigation to the plots.
- **b.** How many different arrangements of the five methods of irrigation are possible in each of the farms?
- **c.** How many different arrangements are possible for the whole study of 10 farms?

Ag. **15.2** Refer to Exercise 15.1. The study was conducted and the yields in pounds of blueberries over a growing season are given in the following table.

	Method of Irrigation					
Farm	Surface	Trickle	Center Point	Lateral	Subirrigation	Farm Mean
1	597	248	391	423	350	401.9
2	636	382	434	461	370	456.6
3	591	348	492	504	460	478.9
4	603	366	468	580	452	493.9
5	649	258	457	449	343	430.9
6	512	321	406	464	340	408.7
7	588	423	466	550	327	470.8
8	689	406	502	526	378	500.0
9	690	400	559	469	419	507.3
10	608	380	469	550	458	493.2
Method Mean	616.3	353.2	464.3	497.6	389.6	464.2

- **a.** Use residual plots to determine if there appear to be a violations in the conditions of normality and equal variance of the residuals.
- **b.** What is the standard error in estimating the mean yield for each of the five methods of irrigation?
- **c.** What is the standard error in estimating the difference in the mean yields of two of the methods of irrigation?
- **d.** Is there significant evidence at the $\alpha = .05$ level that the five methods of irrigation differ in their mean yields?
- **e.** Use a multiple-comparison procedure to determine which pairs of the five methods of irrigation have different means.

Ag. **15.3** Refer to Exercise 15.2. The horticulturist is planning a new study involving modifications to several of the methods of irrigation. In the previous study, it was somewhat cumbersome having blueberry growers implement five different methods of irrigation on their farms, and she wants to know if using the 10 farms as levels of a blocking factor was necessary. If not, she plans to use a single irrigation method on each of n farms.

- **a.** Compute the relative efficiency of the farms as a blocking variable.
- **b.** How many farms would she need in a completely randomized design to have the same precision as was achieved in the randomized block design?

Env. **15.4** Two devices have been proposed to reduce the air pollution resulting from the emission of carbon monoxide (CO) from the exhaust of automobiles. To evaluate the effectiveness of the devices, 48 cars of varying age and mechanical condition were selected for the study. The amount of carbon monoxide in the exhaust (in ppm) was measured prior to installing the device on each of the cars. Because there were considerable differences in the mechanical conditions of the cars, the cars were paired based on the level of CO in their exhaust. The two devices were then randomly assigned to the cars within each pair of cars. Five months after installation, the amount of CO in the exhaust was again measured on each of the cars. The reductions in carbon monoxide from the initial measurements are given here.

Pair	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Before	2.37	3.17	3.07	2.73	3.49	4.35	3.65	3.97	3.21	4.46	3.81	4.55
After	2.51	2.65	2.60	2.40	2.31	2.28	0.94	2.21	3.29	1.92	3.38	2.43
Pair	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Before	4.51	3.03	4.47	3.44	3.52	3.05	3.66	3.81	3.13	3.43	3.26	2.85
After	1.83	2.63	2.31	1.85	2.92	2.26	3.11	1.90	2.50	3.18	3.24	2.16

a. Does there appear to be a difference between the two devices with respect to their ability to reduce the average amount of CO in the exhaust of the cars? Use $\alpha = .05$.

b. Compute the relative efficiency of the randomized complete block design (blocking on car) compared to a completely randomized design in which the 48 cars would have been randomly assigned to the two devices without regard to any pairing. Interpret the value of the relative efficiency.

c. Based on the relative efficiency computed in part (b), would you recommend pairing the cars in future studies?

Env. **15.5** Refer to Exercise 15.4.

a. In Chapter 6, we introduced the paired t test. Analyze the above data using this test statistic.

b. Show that the paired t test is equivalent to the F test from the randomized block AOV by showing that your computed values for the t test and F test satisfy $t^2 = F$. Furthermore, show that the critical values from the t table and F table satisfy the following relationship: $t^2_{.05/2,\,23} = F_{.05,\,1,\,23}$. Therefore, the paired t test and F test from the randomized block AOV must be equivalent.

Psy. **15.6** An industrial psychologist working for a large corporation designs a study to evaluate the effect of background music on the typing efficiency of secretaries. The psychologist selects a random sample of seven secretaries from the secretarial pool. Each subject is exposed to three types of background music: no music, classical music, and hard rock music. The subject is given a standard typing test that combines an assessment of speed with a penalty for typing errors. The particular order of the three experiments is randomized for each of the seven subjects. The results are given here, with a high score indicating a superior performance. This is a special type of randomized complete block design in which a single experimental unit serves as a block and receives all treatments.

	Subject						
Type of Music	**1**	**2**	**3**	**4**	**5**	**6**	**7**
No music	20	17	24	20	22	25	18
Hard rock	20	18	23	18	21	22	19
Classical	24	20	27	22	24	28	16

a. Write a statistical model for this experiment and estimate the parameters in your model.

b. Are there differences in the mean typing efficiencies for the three types of music? Use $\alpha = .05$.

c. Does the additive model for a randomized complete block design appear to be appropriate? (*Hint:* Plot the data as was done in Figure 15.1.)

d. Compute the relative efficiency of the randomized block design compared to a completely randomized design. Interpret this value. Were the blocks effective in reducing the variability in experimental units? Explain.

Psy. **15.7** Refer to Exercise 15.6. Do the model conditions appear to be satisfied?

15.3 Latin Square Design

Ag. **15.8** An experiment compared two different fertilizer placements (broadcast, band) and two different rates of fertilizer flow on watermelon yields. Recent research has shown that broadcast application (scattering over the outer area) of fertilizer is superior to bands of fertilizer applied near the seed for watermelon yields. For this experiment, the investigators wished to compare two nitrogen–phosphorus–potassium fertilizers applied (broadcast and band) at a rate of 160–70–135 pounds per acre and including two brands of micronutrients (A and B). These four combinations were to be studied in a Latin square field plot.

The treatments were randomly assigned according to a Latin square design conducted over a large farm plot, which was divided into rows and columns. A watermelon plant dry weight was obtained for each row–column combination 30 days after the emergence of the plants. The data are shown next.

	Column							
Row	**1**		**2**		**3**		**4**	
1	1	1.75	3	1.43	4	1.28	2	1.66
2	2	1.70	1	1.78	3	1.40	4	1.31
3	4	1.35	2	1.73	1	1.69	3	1.41
4	3	1.45	4	1.36	2	1.65	1	1.73

Treatment 1–broadcast, A Treatment 3–band, A
Treatment 2–broadcast, B Treatment 4–band, B

a. Write an appropriate statistical model for this experiment.

b. Use the data to run an analysis of variance. Give the p-value for each test, and draw conclusions.

Ag. **15.9** Refer to Exercise 15.8.

a. Describe how the four fertilizer placement–rate combinations are randomly assigned to the rows and columns in the farm plot.

b. Compute the relative efficiency of the Latin square design compared to a completely randomized design. Were the row- and column-blocking variables effective in reducing the variability in the responses from the experimental units? Justify your answer.

c. If future studies were to be conducted, would you recommend using both rows and columns as blocking variables? Explain your answer.

Engin. **15.10** A petroleum company was interested in comparing the miles per gallon achieved by four different gasoline blends (A, B, C, and D). Because there can be considerable variability due to differences in driving characteristics and car models, these two extraneous sources of variability were included as blocking variables in the study. The researcher selected four different brands of cars and four different drivers. The drivers and brands of cars were assigned to blends in the manner displayed in the following table. The mileage (in mpg) obtained over each test run was recorded as follows.

	Car Model			
Driver	**1**	**2**	**3**	**4**
1	A(15.5)	B(33.8)	C(13.7)	D(29.2)
2	B(16.3)	C(26.4)	D(19.1)	A(22.5)
3	C(10.5)	D(31.5)	A(17.5)	B(30.1)
4	D(14.0)	A(34.5)	B(19.7)	C(21.6)

a. Write a model for this experimental setting.
b. Estimate the parameters in the model.
c. Conduct an analysis of variance. Use $\alpha = .05$.
d. What conclusions can you draw concerning the best gasoline blend?
e. Compute the relative efficiency of the Latin square design compared to a completely randomized design. Interpret this value. Were the blocking variables effective in reducing the variability in experimental units? Explain.
f. If future studies were to be conducted, would you recommend using both car model and driver as blocking variables? Explain.

Engin. **15.11** Refer to Exercise 15.10.
a. Do the model conditions appear to be satisfied for this set of data? Explain.
b. If the model conditions appear to be violated, suggest an alternative method of analysis.

15.4 Factorial Treatment Structure in a Randomized Complete Block Design

Med. **15.12** A psychologist is designing a study to evaluate three new treatments for a behaviorial problem in children. The psychologist will include a second factor, which will classify the subjects according to four levels of socioeconomic status. There are 30 children available for each level of socioeconomic level, which will provide 10 replications of each of the treatments by socioeconomic combinations. At the end of the treatment period, the children will be assessed and assigned a score reflecting the degree of improvement in their behavior. There are five trained evaluators who will assign the scores to the children. The psychologist knows from past studies that some evaluators tend to assign uniformly higher scores than other evaluators, and, hence, he wants to be able to control for the evaluator effect in the analysis of the treatment–socioeconomic status effect.
a. Display how you would randomly assign the children to the 12 treatment–socioeconomic status combinations.
b. Provide an analysis of variance table for this experiment (source of variation and degrees of freedom).

Ag. **15.13** An entomologist employed by a chemical company is planning a study to evaluate two new chemicals that are potential agents for eliminating fire ants. The chemicals will be evaluated at three different dose levels under four different environmental conditions. One hundred ants will be exposed to each of the combinations of a chemical, dose level, and environmental condition, and the number of surviving ants after 3 hours of exposure will be recorded. It is well documented in the literature that there is large variability in the degree of tolerance of fire ants to various chemicals previously used as insecticides. Thus, the company's statistician recommended that five colonies of ants be used in the study. There are thousands of fire ants per colony.
a. Display how you would randomly assign the groups of 100 ants to the various combinations of chemical–dose–environmental condition.
b. Provide an analysis of variance table for this experiment (source of variation and degrees of freedom).

Gov. **15.14** The transportation research division of a northern state is examining the amount of road damage associated with various methods used to clear snow and ice from the roadways. The

division engineers have selected two levels of each of the following substances that are applied to the roadways: sodium chloride, calcium chloride, and sand. The response variable measured on each of the treated roads is the number of new cracks per mile of roadway. Because traffic volume is highly variable and could impact the response variable, the engineers decide to use a randomized block design with the traffic volume during the previous winter as the blocking factor. Each of the six treatments is randomly assigned to five roadways. The data are given here.

	Sodium Chloride		Calcium Chloride		Sand	
Roadway	**Low**	**High**	**Low**	**High**	**Low**	**High**
1	37	49	43	47	27	33
2	39	50	42	48	27	31
3	48	52	47	50	36	37
4	44	57	45	54	34	37
5	54	68	56	63	45	44

a. Write a statistical model for this experiment.
b. Use a profile plot to display the interaction between treatment and level.
c. Perform appropriate F tests, and draw conclusions from these tests concerning the effect of treatment and level on the mean number of cracks.
d. Use a normal probability plot and a plot of the residuals to determine if there are violations in the appropriate conditions for validly drawing conclusions from the F tests.

Gov. **15.15** Refer to Exercise 15.14.

a. Describe how the treatments would be randomly assigned to the roadways.
b. Compute the relative efficiency of the randomized block design compared to a completely randomized design. Was the blocking of the roadways based on traffic volume effective in reducing the variability in the counts of number of cracks? Explain.
c. If this study was repeated during the next winter, would you recommend that traffic volume be used to block the roadways, or would it be more efficient to design the study as a completely randomized design?

Ag. **15.16** An agricultural experiment station is investigating the appropriate planting density for three commercial varieties of tomatoes: celebrity, sunbeam, and trust. The researcher decides to examine the effects of four planting densities: 5, 20, 35, and 50 thousand plants per hectare. The experiment station has three large fields that would be appropriate for the study. At each of the fields, 12 plots are prepared, and the 12 treatments are randomly assigned to the plots. A separate randomization is done at each of the three fields. The yield, in tons, from the 36 one hectare plots are given here.

	Variety											
	Celebrity				Sunbeam				Trust			
	Density				Density				Density			
Field	**5k**	**20k**	**35k**	**50k**	**5k**	**20k**	**35k**	**50k**	**5k**	**20k**	**35k**	**50k**
1	32.5	39.9	42.5	38.2	32.2	43.2	47.6	43.5	49.9	59.0	66.3	58.3
2	33.4	47.2	44.5	43.5	33.4	51.3	52.2	44.1	60.8	66.1	70.7	60.6
3	41.1	48.7	53.5	48.4	41.8	51.2	55.9	55.9	60.8	67.6	73.2	67.8

a. Identify the design, and write a statistical model for this experiment.
b. Use a profile plot to display the level of interaction between treatment and level.
c. Perform appropriate F tests, and draw conclusions from these tests concerning the effect of variety and planting density on the mean yield of the tomato plants.
d. Use a normal probability plot and a plot of the residuals to determine if there are violations in the appropriate conditions for validly drawing conclusions from the F tests.

Ag. **15.17** Refer to Exercise 15.16.

a. Describe how the varieties of plants and planting densities would be randomly assigned to the plots of land.

b. Compute the relative efficiency of the randomized block design compared to a completely randomized design. Do you think it was necessary for the researchers to block on fields? Explain.

c. During the summer months when the experiment was conducted, it was unusually hot, and the researcher decides to repeat the experiment during the next growing season. The researcher would like to use the same three fields, but this time he would like to plant celebrity plants on field 1, sunbeam on field 2, and trust on field 3. Explain to the researcher why this design may not be appropriate.

Ag. **15.18** Refer to Exercise 15.16.

a. Which pairs of varieties appear to have significantly different mean yields at the $\alpha = .05$ level?

b. Which pairs of planting densities appear to have significantly different mean yields at the $\alpha = .05$ level?

c. Which variety appears to produce the largest mean yield?

d. Which planting density appears to produce the largest mean yield?

e. Explain what aspect of your model allows you to answer part (c) without referring to planting density?

15.5 A Nonparametric Alternative—Friedman's Test

15.19 Refer to Exercise 15.2.

a. What are the conditions under which it is appropriate to use the Friedman test in comparing the mean yields from the five irrigation methods?

b. Use the Friedman test to determine if there is significant evidence of a difference in the mean yields for the five irrigation methods.

c. Compare the conclusions obtained from the Friedman test to the conclusions obtained from the AOV F test.

d. Explain why the conclusions should be different (or the same).

15.20 Refer to Exercise 15.14.

a. Use the Friedman test to determine if there is significant evidence of a difference in the mean number of counts for the six potential treatments for removing ice and snow from the roadway.

b. Compare the conclusions obtained from the Friedman test to the conclusions obtained from the AOV F test.

15.21 Refer to Exercise 15.16.

a. Use the Friedman test to determine if there is significant evidence of a difference in the mean yields for the 12 combinations of variety–planting density.

b. Compare the conclusions obtained from the Friedman test to the conclusions obtained from the AOV F test.

Supplementary Exercises

Sci. **15.22** An experiment compares four different mixtures of the components oxidizer, binder, and fuel used in the manufacturing of rocket propellant. The four mixtures under test, corresponding to settings of the mixture proportions for oxide, are shown here.

Mixture	Oxidizer	Binder	Fuel
1	.4	.4	.2
2	.4	.2	.4
3	.6	.2	.2
4	.5	.3	.2

To compare the four mixtures, five different samples of propellant are prepared from each mixture and readied for testing. Each of five investigators is randomly assigned one sample of each of the four mixtures and asked to measure the propellant thrust. These data are summarized next.

	Investigator				
Mixture	**1**	**2**	**3**	**4**	**5**
1	2,340	2,355	2,362	2,350	2,348
2	2,658	2,650	2,665	2,640	2,653
3	2,449	2,458	2,432	2,437	2,445
4	2,403	2,410	2,418	2,397	2,405

a. Identify the blocks and treatments for this experimental design.
b. Indicate the method of randomization.
c. Why would this design be preferable to a completely randomized design?

Sci. **15.23** Refer to Exercise 15.22.
a. Write a model for this experimental setting.
b. Estimate the parameters in the model.
c. Display a complete analysis of variance table. Use $\alpha = .05$.
d. What conclusions can you draw concerning the best mixture from the four tested? (*Note:* The higher the response value, the better the rocket propellant's thrust.)
e. Compute the relative efficiency of the randomized block design compared to a completely randomized design. Interpret this value. Were the blocks effective in reducing the variability in experimental units? Explain.

Engin. **15.24** A quality control engineer is considering implementing a workshop to instruct workers on the principles of total quality management (TQM). The program would be quite expensive to implement across the whole corporation; hence, the engineer has designed a study to evaluate which of four types of workshops would be most effective. The response variable will be the increase in productivity of the worker after participating in the workshop. Since the effectiveness of the workshop may depend on the worker's preconceived attitude concerning TQM, the workers are given an examination to determine their attitudes prior to taking the workshop. Their attitudes are classified into five groups. There are four workers in each group, and the type of workshop is randomly assigned to the workers within each group. The increases in productivity are given here.

	Attitude					
Type of Workshop	**1**	**2**	**3**	**4**	**5**	**Mean**
A	33	38	39	42	62	42.8
B	35	37	43	47	71	46.6
C	40	42	45	52	74	50.6
D	54	50	55	62	84	61.0
Mean	40.5	41.75	45.5	50.75	72.75	50.25

a. Write a statistical model for this experiment, and estimate the parameters in your model.
b. Are there differences in the mean increases in productivity for the four types of workshops? Use $\alpha = .05$.
c. Does the additive model for a randomized complete block design appear to be appropriate? (*Hint:* Plot the data as in Figure 15.1.)
d. Compute the relative efficiency of the randomized block design compared to a completely randomized design. Interpret this value. Were the blocks effective in reducing the variability in experimental units? Explain.

Engin. **15.25** Refer to Exercise 15.24. Based on the residuals from the fitted model, do the model conditions appear to be satisfied?

Engin. **15.26** An experimenter is interested in examining the bond strength of a new adhesive product prepared under three different temperature settings (280°F, 300°F, and 320°F) and four different pressure settings (100, 150, 200, and 250 psi). The experimenter will prepare a sufficient amount of the adhesive so that each temperature–pressure setting combination is tested on three samples of the adhesive. Suppose that the experimenter can test only 12 samples per day and that the conditions in the laboratory are somewhat variable from day to day. Describe an experimental design that takes into account the day-to-day variation in the laboratory. Include a diagram that displays the assignment of the temperature–pressure setting combinations to adhesive samples.

Edu. **15.27** A study was conducted to study the impact of child abuse on performance in school. Three categories of child abuse were defined as follows:

Abused child—a child who is physically abused.

Neglected child—a child receiving inadequate care.

Nonabuse—a child receiving normal care and not physically abused.

The researchers randomly selected 30 boys and 30 girls from each of the three categories using the records of the state child-welfare agency for the abused and neglected children and the records of a local school for the nonabused children. The scores on a standard grade-level assessment test of reading, mathematics, and general science were recorded for all the selected children.

a. Suppose the children were all in the seventh grade. Identify the design.
b. Suppose the children were equally divided among the third, fifth, and seventh grades. Identify the design.

Gov. **15.28** The city manager of a large midwestern city was negotiating with the three unions that represented the police, firefighters, and building inspectors over the salaries for these groups of employees. The three unions claimed that the starting salaries were substantially different among the three groups, whereas in most cities there was not a significant difference in starting salaries among the three groups. To obtain information on starting salaries across the nation, the city manager decided to randomly select one city in each of eight geographical regions. The starting yearly salaries (in thousands of dollars) were obtained for each of the three groups in each of the eight regions. The data appear here.

Region	1	2	3	4	5	6	7	8	Mean
Police	32.3	33.2	30.8	30.5	30.1	30.2	28.4	27.9	30.42
Firefighters	31.9	32.8	31.6	31.2	30.8	30.6	28.7	27.5	30.64
Inspectors	27.9	27.8	26.5	26.8	26.4	26.8	25.3	25.9	26.68
Region mean	30.7	31.3	29.6	29.5	29.1	29.2	27.5	27.1	29.25

a. Write a model for this study, identifying all the terms in the model.
b. Do the data suggest a difference in mean starting salaries for the three groups of employees? Use $\alpha = .05$.
c. Give the level of significance for your test.
d. Which pairs of jobs types have significantly different starting salaries?

Gov. **15.29** Refer to Exercise 15.28.

a. Plot the data in a profile plot with factors job type and region. Does there appear to be an interaction between the two factors? If there was an interaction, would you be able to test for it using the given data? If not, why not?
b. Did the geographical region variable increase the efficiency of the design over conducting the study as a completely randomized design in which the city manager would have randomly selected eight cities regardless of their location?
c. Identify additional sources of variability that may need to be included in future studies.

Ag. **15.30** Refer to Exercise 14.23. In the description of this experiment, the researchers failed to note that the experiment in fact had been conducted at four different orange groves, which were

located in different states. Grove 1 had a soil pH of 4.0, grove 2 had a soil pH of 5.0, grove 3 had a soil pH of 6.0, and grove 4 had a soil pH of 7.0. At each of the groves, three trees were randomly assigned to one of the calcium levels: 100, 200, or 300 pounds per acre. The data are given here.

		Calcium		
Grove	**pH Value**	**100**	**200**	**300**
1	4.0	5.2, 5.9, 6.3	7.4, 7.0, 7.6	6.3, 6.7, 6.1
2	5.0	7.1, 7.4, 7.5	7.4, 7.3, 7.1	7.3, 7.5, 7.2
3	6.0	7.6, 7.2, 7.4	7.6, 7.5, 7.8	7.2, 7.3, 7.0
4	7.0	7.2, 7.5, 7.2	7.4, 7.0, 6.9	6.8, 6.6, 6.4

a. How would this new information alter the conclusions reached in Exercise 14.23 concerning the effect of soil pH and calcium on the mean increases in tree diameter?
b. Design a new experiment in which the effects of soil pH and calcium on the mean increases in tree diameter could be validly evaluated. All four groves must be used in your design, along with the four levels of pH and three levels of calcium.

Bus. **15.31** A food-processing plant has tested several different formulations of a new breakfast drink. Each of six panels rated the 12 different formulations obtained from combining one of three levels of sweetness, one of two levels of caloric content, and one of two colors. The mean ratings are given in the following table.

	Color			
	1		2	
Sweetness	**Caloric Level**		**Caloric Level**	
Level	**1**	**2**	**1**	**2**
1	59.5	42.5	54.5	40.1
2	66.8	49.6	64.7	50.1
3	52.0	39.3	35.1	30.2

a. Identify the design.
b. Write an appropriate model.
c. Give the analysis of variance table for this design.

Bus. **15.32** The following AOV table was computed for the experimental design described in Exercise 15.31. What is missing from the table?

Source	**SS**	**df**	**MS**	***F*-Value**	**Pr > *F***
Main effects					
A	4,149.55556	2	2,074.76389	75.51	.0001
B	624.22222	1	624.22222	22.72	.0001
C	3,200.00000	1	3,200.00000	116.46	.0001
Interactions					
AB	488.52778	2	244.26389	8.89	.0004
AC	203.08333	2	101.54167	3.70	.0307
BC	80.22222	1	80.22222	2.92	.0927
ABC	24.19444	2	12.09722	.44	.6459
Error	1,648.66667	60	27.47778		

Engin. **15.33** Three dye formulas for a certain synthetic fiber are under consideration by a textile manufacturer who wishes to know whether the three are in fact different in quality. To aid in this decision, the manufacturer conducts an experiment in which five specimens of fabric are cut into

thirds, and one third is randomly assigned to be dyed by each of the three dyes. Each piece of fabric is later graded and assigned a score measuring the quality of the dye. The results are as follows.

	Fabric Specimen				
Dye	**1**	**2**	**3**	**4**	**5**
A	74	78	76	82	77
B	81	86	90	93	73
C	95	99	90	87	93

a. Identify the design.
b. Run an analysis of variance, and draw conclusions about the dyes. Use $\alpha = .05$.
c. Give a measure of the efficiency of this design compared to one not blocking on fabric specimens.

Psy. **15.34** An experiment tested the effect of music on factory workers' production. Four music programs (A, B, C, and D) were compared with no music (E). Each program was played for an entire day, and five replications for each program were desired. The length of the experiment was thus 5 weeks. To control for variation in week and day of the week, a Latin square design was adopted for the 25 days of the experiment. Each program was played once on each day of the week and once each week.

Week	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
1	133 (E)	139 (B)	140 (C)	140 (D)	145 (A)
2	139 (A)	136 (E)	141 (B)	143 (C)	146 (D)
3	138 (B)	139 (D)	140 (E)	139 (A)	142 (C)
4	137 (C)	140 (A)	136 (D)	129 (E)	132 (B)
5	142 (D)	143 (C)	142 (A)	144 (B)	132 (E)

a. Does there appear to be a difference in mean workers' production totals among the five types of music? Use $\alpha = .05$.
b. If there is a difference in mean workers' production totals, which of the four music programs appear to be associated with higher mean workers' production totals in comparison to no music?

Ag. **15.35** The yields of wheat (in pounds) are shown here for five farms. Five plots are selected based on their soil fertility at each farm, with the most fertile plots designated as 1. The treatment (fertilizer) applied to each plot is shown in parentheses.

	Plot				
Farm	**1**	**2**	**3**	**4**	**5**
1	(D) 10.3	(E) 8.6	(A) 6.7	(C) 7.6	(B) 5.8
2	(E) 8.8	(B) 6.7	(C) 6.7	(A) 4.8	(D) 6.0
3	(A) 6.3	(C) 8.3	(B) 6.8	(D) 8.0	(E) 8.8
4	(C) 8.9	(D) 7.4	(E) 8.2	(B) 6.2	(A) 4.4
5	(B) 7.3	(A) 4.4	(D) 7.7	(E) 6.8	(C) 6.7

a. Identify the design.
b. Do an analysis of variance, and draw conclusions concerning the five fertilizers. Use $\alpha = .01$.

Ag. **15.36** Refer to Exercise 15.35. Run a multiple-comparison procedure to make all pairwise comparisons of the treatment means.

Med. **15.37** A medical researcher designed an experiment to study the impact of three exercise regimens (30, 60, and 90 minutes per week) on the total blood cholesterol level in active adult males.

The researcher was concerned that the effect of the type of exercise program on cholesterol might also depend on the age of the individual. Nine participants in each of three age groups (A1: 20–29, A2: 30–39, A3: 40–49) were obtained from five fitness centers. The total blood cholesterol level of the participants were measured both prior to the start of the study and after 6 months on the exercise regimens. The reductions in total blood cholesterol level are given in the following table.

	Exercise Regimen								
	90 min.			60 min.			30 min.		
Center	A1	A2	A3	A1	A2	A3	A1	A2	A3
1	82	49	54	52	57	67	39	46	–3
1	50	41	17	32	7	7	30	55	–3
1	31	36	47	26	17	29	–9	–28	18
2	43	60	51	3	25	–8	4	32	23
2	34	24	3	64	–13	–14	–26	7	–6
2	–18	14	–23	34	30	30	45	53	2
3	38	41	15	–3	15	7	17	3	23
3	–6	65	0	–4	9	–10	56	–9	3
3	30	18	23	–15	24	0	–30	–7	–16
4	38	–7	51	–2	35	–12	–13	19	15
4	–7	7	–3	21	–13	–11	–14	–18	–32
4	–30	–36	–36	–44	–9	10	15	–4	14
5	–3	18	35	3	–13	–22	26	3	–28
5	–3	–4	–55	–37	–1	–30	11	–7	–19
5	–37	4	–12	8	–20	–43	–38	–22	–32

a. Identify the design by name.
b. Write a model for this study, identifying all the terms in the model.
c. Do the data support the research hypothesis that the mean reduction in cholesterol increases with an increase in exercise? Use $\alpha = .05$ in reaching your conclusion.
d. Is your answer in part (c) consistent across all three age groups? Support your answer with a p-value from an appropriate test of hypotheses.
e. Assume that the effectiveness of the three exercise regimens differed for the three age groups. Group the three exercise regimens separately for each age group using an overall Type I error rate of .05.

Med. **15.38** Refer to Exercise 15.37.
a. If this experiment was conducted again, would you recommend including the factor associated with center in your model and analysis?
b. What was the relative efficiency of the center factor in the analysis of the data in Exercise 15.37?

Med. **15.39** Refer to Exercise 15.37.
a. Does a residual analysis support the conditions necessary to conduct your tests in Exercise 15.37?
b. Conduct an analysis of the data, assuming that the normality condition does not hold, using a rank-based procedure.
c. Compare your conclusions about exercise regimens and age groups reached using the rank-based procedure to your conclusions reached using the normal-based procedures. Which set of conclusions would be more easily supported using the given data?

Engin. **15.40** ***Mason, Gunst, and Hess (2003)*** describe the following study. A traffic engineer designs a study to compare the total unused red-light times for five methods of traffic-light signaling (A, B, C, D, and E). The engineer randomly selects five intersections in a major city and five time periods

spaced across the day. The following table contains the unused red-light time in minutes, with the letter in parentheses indicating the method of signaling.

	Period of Day				
Intersection	**1**	**2**	**3**	**4**	**5**
1	15.2(A)	33.8(B)	13.5(C)	27.4(D)	29.1(E)
2	16.5(B)	26.5(C)	19.2(D)	25.8(E)	22.7(A)
3	12.1(C)	31.4(D)	17.0(E)	31.5(A)	30.2(B)
4	10.7(D)	34.2(E)	19.5(A)	27.2(B)	21.6(C)
5	14.6(E)	31.7(A)	16.7(B)	26.3(C)	23.8(D)

a. Identify the design by name.
b. Write a model for this study, identifying all the terms in the model.
c. Describe how randomization could be conducted in this study.
d. Is there significant evidence of a difference in the mean unused red-light times for the five signaling methods? Use $\alpha = .05$.
e. Group the five intersections on the basis of their mean unused red-light times.

Engin. **15.41** Refer to Exercise 15.40.
a. What was the relative efficiency of the period of day factor in the analysis of the data in Exercise 15.40?
b. What was the relative efficiency of the intersection factor in the analysis of the data in Exercise 15.40?
c. In future traffic studies, would you recommend including factors to control for the variation in intersection and/or period of day?

Engin. **15.42** Refer to Exercise 15.41. Based on a residuals analysis, do the necessary conditions for conducting the test of hypotheses appear to be valid?

Edu. **15.43** An educational researcher designs a study to evaluate the effect of providing students with a laptop containing supplemental material to assist them in learning specified mathematical concepts. The researcher wants to also evaluate the effect of grade level of the students and their mathematical ability on the benefit of using the supplemental materials. The principals of two schools, one junior high and one high school, agree to participate in the study. Within each school, 12 classrooms are selected, with 2 classrooms randomly assigned to each of the combinations of two factors: supplemental materials (yes or no) and student math scores in the previous school years (low, medium, and high). Twenty students in each of the 24 classrooms are given a test to evaluate their mathematical proficiency both at the beginning and at the end of the semester in which the study was conducted. The difference in the two test scores will be used as the response variable to measure whether the supplemental materials provided a benefit in learning mathematical concepts. The mean responses of the students in the 24 classrooms are given in the following table.

School	Supplements	Math Ability	Response	School	Supplements	Math Ability	Response
JunHigh	Yes	Low	22.3	HighSch	Yes	Low	24.2
JunHigh	Yes	Low	14.7	HighSch	Yes	Low	35.6
JunHigh	Yes	Med	29.1	HighSch	Yes	Med	38.9
JunHigh	Yes	Med	31.8	HighSch	Yes	Med	49.5
JunHigh	Yes	Hgh	29.6	HighSch	Yes	Hgh	44.7
JunHigh	Yes	Hgh	42.3	HighSch	Yes	Hgh	54.3
JunHigh	No	Low	12.9	HighSch	No	Low	11.5
JunHigh	No	Low	17.3	HighSch	No	Low	24.3
JunHigh	No	Med	16.8	HighSch	No	Med	34.1
JunHigh	No	Med	22.7	HighSch	No	Med	28.4
JunHigh	No	Hgh	27.1	HighSch	No	Hgh	34.2
JunHigh	No	Hgh	25.3	HighSch	No	Hgh	31.0

a. The researcher analyzed the data as a completely randomized experiment with two replications of the complete crossing of the three factors: type of school (junior high or high school), supplemental materials (yes or no), and math ability of the students (low, medium, or high). If possible, test for the main effects, two-way interactions, and three-way interaction of the three factors at the $\alpha = .05$ level.

b. If you determined that it was not possible to conduct all the tests requested in part (a), modify the analysis so that a complete analysis can be conducted on two of the three factors.

CHAPTER 16

The Analysis of Covariance

16.1 Introduction and Abstract of Research Study

In some experiments, the experimental units are nonhomogeneous, or there is variation in the experimental conditions that is not due to the treatments. For example, a study is designed to evaluate different methods of teaching reading to 8-year-old children. The response variable is final scores of the children after participating in the reading program. However, the children participating in the study will have different reading abilities prior to entering the program. Also, there will be many factors outside the school that may have an influence on the reading score of the children, such as socioeconomic variables associated with a child's family. The variables that describe the differences in experimental units or experimental conditions are called **covariates.** The analysis of covariance is a method by which the influence of the covariates on the treatment means is reduced. This will often result in increased precision for parameter estimates and increased power for tests of hypotheses.

covariates

In Chapter 15, we addressed this problem through the use of randomized complete block and Latin square designs. The experimental units were grouped into blocks of experimental units, which provided for greater homogeneity of the experimental units within each block than was present in the collection of experimental units as a whole. Thus, we achieved a reduction in the variation of the responses due to factors other than the treatments.

In many experiments, it may be difficult or impossible to block the experimental units. The characteristics that differentiate the experimental units may not be known prior to running the experiment, or the variables that affect the response may not surface until after the experiments have started. In some cases, there may

be too few experimental units in each block to examine all the treatments. Several examples of these types of experiments include the following:

- A clinical trial is run to evaluate the several traditional methods for treating chronic pain and some new alternative approaches. The patients included in the trial would have different levels of pain depending on the length of time they have been inflicted with the syndrome, their ages, their physical conditions, and many other factors that can affect the performance of the treatment. Researchers could block on several of these factors, but the influence of the other covariates may have an undue influence on the outcome of the trial.
- The aerial application of insecticides to control fire ants is proposed for large pasturelands in Texas. There are a number of possible methods for applying the insecticide to the pastures. Because the EPA is concerned about the spray drifting off the target areas, a study is designed to evaluate the accuracy of the spraying techniques. The amount of the insecticide, y, landing within the target areas is recorded for each of the four methods of applying the insecticide. The testing is to be conducted only on those days in which there is little or no wind. However, in Texas there are always wind gusts that may affect the accuracy of the spraying. Thus, an important covariate is the wind speed at the target area during the spraying.
- A fiber-optic cable manufacturer is investigating three new machines used in coating the cable. The response of interest is the tensile strength, y, of the cable after the coating is applied. Although the coating is set at a uniform thickness of 1.5 mm, there is some variation in thickness along the length of a 100-meter cable. This variation in thickness may affect the tensile strength of the cable. The testing is conducted in a laboratory with a constant temperature. The experiments are run over a 5-day period of time. Because there are some environmental and technician differences in the laboratory from day to day, the researchers decide to block on day and to record the thickness of the coating at the break point in the cable. Thus, both a blocking variable and a covariate will be involved in the experiment.

The following research study involves an experiment in which the measured response is related not only to the assigned treatment but also to a covariate, which was measured on the experimental unit during the study.

Abstract of Research Study: Evaluation of Cool-Season Grasses for Putting Greens

A problem confronting greenskeepers on golf courses is the prevalence of viral diseases, which damage putting greens. The diseases are particularly dangerous during the early spring when the weather is cool and wet and the grasses on the greens have not completely recovered from winter dormancy. Several new cultivars of turfgrass for use on golf course greens have been developed. These cultivars are resistant to the type of viral diseases that are of concern to the greenskeepers. Prior to adopting the grasses for use on golf course greens, it was necessary to evaluate the cultivars with respect to their appropriateness for use on the putting surfaces. From previous studies, three cultivars (C_1, C_2, and C_3) were found to have the

TABLE 16.1
Green speed of three cultivars

	C_1		C_2		C_3	
Region	**Humidity**	**Speed**	**Humidity**	**Speed**	**Humidity**	**Speed**
1	31.60	7.56	29.42	8.88	89.60	8.20
2	54.12	7.41	44.44	8.20	37.17	9.15
3	42.34	7.64	84.38	7.20	37.32	9.24
4	53.82	6.81	88.42	7.12	89.21	8.31
5	86.70	6.86	71.33	8.16	58.57	9.42
6	76.27	6.86	45.50	8.68	66.68	9.26
7	68.66	7.22	66.79	8.25	82.78	8.93
8	47.27	7.64	58.34	8.22	29.52	9.89

greatest resistance to the early spring viral diseases. The researchers determined from discussions with golf course superintendents that the performance measure of greatest interest was the speed that a ball rolls on the green after being struck by a putter. The United States Golf Association (USGA) has developed a device called the Stimpmeter to evaluate the speed of the greens. The Stimpmeter is a 36-inch extruded aluminum bar with a grooved runway on one side. A notch in the runway is used to support a golf ball until one end of the Stimpmeter is lifted to an angle of roughly 20 degrees. The average distance the golf ball travels after two opposing rolls down the Stimpmeter is referred to as the speed of the green. The farther the ball rolls, the faster the green. Important factors that affect speed are the length of the grass, hardness of the surface, and slope of the surface.

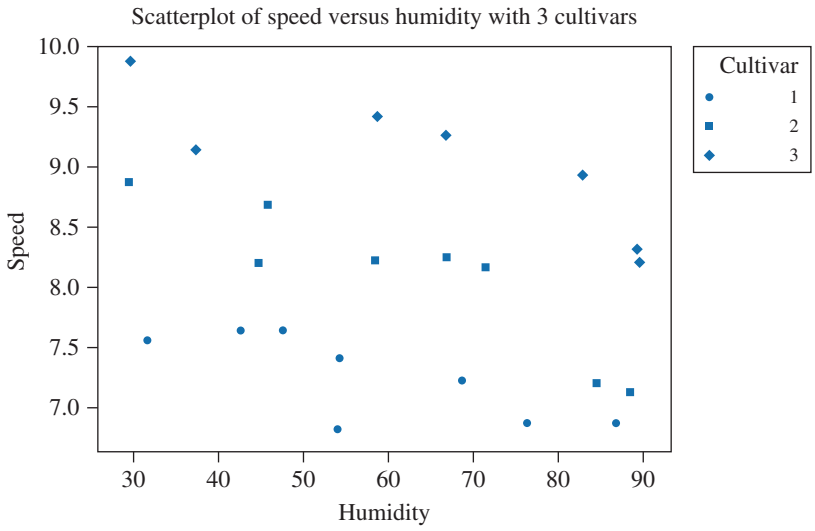
The researchers decided to study eight different regions of the country. In each region, a golf course was selected, and three putting greens were constructed. The three greens had the same soil composition and slope. The three cultivars were randomly assigned to a single green at each of the eight golf courses. Thus, the factors affecting green speed that are associated with geographical location were controlled through the use of blocking. A factor that was considered to be important but that the researchers were not able to control was the humidity during the testing period. Thus, it was decided to record humidity and use it as a covariate. The measurements of green speed (in feet) and humidity at the eight locations are given in Table 16.1.

The speed measurement for each of the greens is plotted in Figure 16.1 versus the humidity reading during the testing period. The plotted points suggest a negative relationship between speed and humidity level, with the relationship similar for all three cultivars. However, cultivar C_3 appears to yield a uniformly greater speed value than the other two cultivars.

In Section 16.5, we will present a model that will enable us to adjust the speed readings for both the region of the country in which the greens were located and the humidity during the time in which the tests were conducted. The three cultivars will then be compared using the adjusted mean speed readings.

Since the analysis of covariance combines features of the analysis of variance and regression analysis, we will make use of a general linear model formulation for the analysis of this type of data. By referring to and building on our work with general linear models in preceding chapters, we can more easily understand the blending of analysis of variance with regression modeling. We begin our presentation with a single covariate in a completely randomized design.

FIGURE 16.1
Speed of golf greens for three cultivars with humidity readings



16.2 A Completely Randomized Design with One Covariate

A completely randomized design is used to compare t population means. To do this, we obtain a random sample of n_i observations on the variable y in the ith population ($i = 1, 2, \ldots, t$). Now, in addition to measuring the response variable y on each experimental unit, we measure a second variable, x, often called a *covariable* or a *covariate*. For example, in studying the effects of different methods of reinforcement on the reading achievement levels of 8-year-old children, we could measure not only the final achievement level y for each child but also the prestudy reading performance level x. Ultimately, we would want to make comparisons among the different methods while taking into account information on both y and x.

Note that x can be thought of as an independent variable, but unlike most situations discussed in previous chapters, here we cannot control the value of x (as we controlled settings of temperature or pressure) prior to observing the variable. In spite of this, we may still write a model for the completely randomized design, treating the covariate as an independent variable.

We will examine an experiment comparing $t = 3$ treatments from a completely randomized experiment with one covariate to illustrate the analysis of covariance procedures.

EXAMPLE 16.1

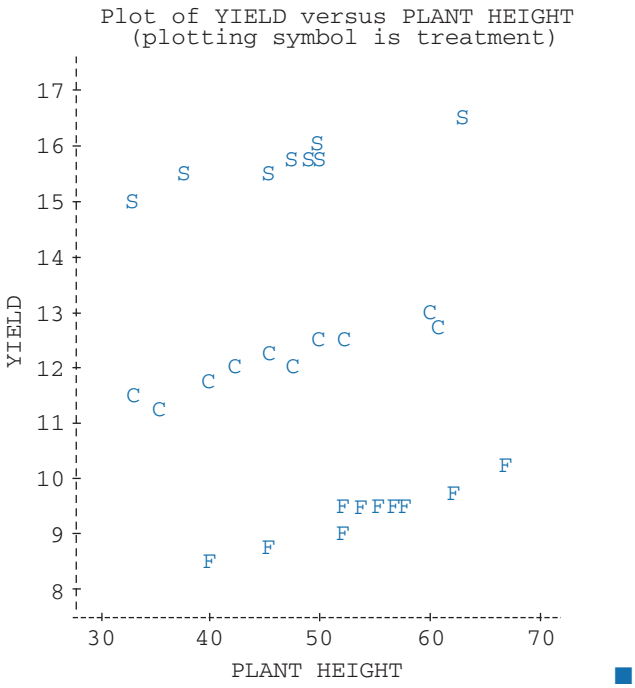
In this study, the effects of two treatments, a slow-release fertilizer (S) and a fast-release fertilizer (F), on seed yield (grams) of peanut plants were compared with a control (C), a standard fertilizer. Ten replications of each treatment were to be grown in a greenhouse study. When setting up the experiment, the researcher recognized that the 30 peanut plants were not exactly at the same level of development or health. Consequently, the researcher recorded the height (cm) of the plant, a measure of plant development and health, at the start of the experiment, as shown in Table 16.2. Plot seed yield versus plant height for the 30 peanut plants.

TABLE 16.2
Peanut plant growth data

Control (C)		Slow Release (S)		Fast Release (F)	
Yield	**Height**	**Yield**	**Height**	**Yield**	**Height**
12.2	45	16.6	63	9.5	52
12.4	52	15.8	50	9.5	54
11.9	42	16.5	63	9.6	58
11.3	35	15.0	33	8.8	45
11.8	40	15.4	38	9.5	57
12.1	48	15.6	45	9.8	62
13.1	60	15.8	50	9.1	52
12.7	61	15.8	48	10.3	67
12.4	50	16.0	50	9.5	55
11.4	33	15.8	49	8.5	40

Solution A plot of the yields for each treatment is shown in Figure 16.2, with the covariate, plant height, given on the horizontal axis.

FIGURE 16.2
Seed yield for three treatments with the covariate, plant height



The experiment described in Example 16.1 was conducted using a completely randomized design with three treatment groups and a single covariate. If we assume a straight-line relationship between seed yield, y_{ij}, and the covariate, plant height, x_{ij}, the model for the completely randomized design with a single covariate is given by

$$y_{ij} = \mu_i + \beta_1(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij}$$

or

$$y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \varepsilon_{ij}$$

with $i = 1, 2, \ldots, t$ and $j = 1, 2, \ldots, n$, where μ_i is the ith treatment mean, β_1 is the slope of the regression of y_{ij} on x_{ij}, β_0 is intercept of the regression of y_{ij} on x_{ij}, τ_i is the ith treatment effect, and ε_{ij} are random independent, normally distributed experimental errors with mean 0 and variance σ_ε^2. The other major conditions imposed on the model in an analysis of covariance are as follows:

1. The relationship between the response y and the covariate x is linear.
2. The regression coefficient β_1 is the same for all treatments.
3. The treatments do not have an effect on the covariate, x_{ij}.

adjusted treatment means

The analysis of covariance involves fitting a number of models to the response variable, y. First, we evaluate whether the covariate, x, provides a significant reduction in the experimental error. If the reduction is significant, then we replace the observed treatment means, $\bar{y}_{i.}$, with estimated **adjusted treatment means,** $\hat{\mu}_{\text{Adj},i}$, which are adjusted for the effect of the covariate on the response variable. Inferences about the treatment differences are then made on the basis of the adjusted means and not the observed means.

We will formulate the required models needed in the analysis of covariance. The model relating y_{ij} to the t treatments and the covariate can be written in the form of an analysis of variance model and then reformulated in regression form.

$$\text{Full model:} \quad y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \varepsilon_{ij}$$

Next, we will formulate two reduced models, one without the covariate and then one without treatment differences but with the covariate.

$$\text{Reduced model I:} \quad y_{ij} = \beta_0 + \tau_i + \varepsilon_{ij}$$

$$\text{Reduced model II:} \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$$

These three models also can be written in the form of the regression (general linear) models of Chapter 12. We make this transition to regression models because it facilitates analysis using various statistical software packages.

$$\text{Full model:} \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_t x_t + \varepsilon$$

where

x_1 = covariate

$x_2 = 1$ if treatment 2 is used $\quad x_2 = 0$ otherwise

$x_3 = 1$ if treatment 3 is used $\quad x_3 = 0$ otherwise

...

$x_t = 1$ if treatment t is used $\quad x_t = 0$ otherwise

It is helpful with these models to refer to a table of expected values, μ_i, as shown in Table 16.3, based on the full model. Note that the treatments have the same slope (β_1) but different intercepts, $\beta_0 + \beta_i, i = 2, \ldots, t$.

TABLE 16.3
Expected values for the full model

Treatment	Expected Value
1	$\mu_1 = \beta_0 + \beta_1 x_1$
2	$\mu_2 = (\beta_0 + \beta_2) + \beta_1 x_1$
⋮	⋮
t	$\mu_t = (\beta_0 + \beta_t) + \beta_1 x_1$

We next fit a reduced model in which the covariate is removed in order to determine the influence of the covariate.

$$\text{Reduced model I:} \quad y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_t x_t + \varepsilon$$

A second reduced model is fit in which the treatment effects are removed but the covariate remains in the model.

$$\text{Reduced model II:} \quad y = \beta_0 + \beta_1 x_1 + \varepsilon$$

From each of these models, we obtain the sum of squares error, which we will denote as follows:

$$\begin{aligned} \text{SSE}_{\text{F}} &= \text{sum of squares error from the full model} \\ \text{SSE}_{\text{RI}} &= \text{sum of squares error from reduced model I} \\ \text{SSE}_{\text{RII}} &= \text{sum of squares error from reduced model II} \end{aligned}$$

The significance of the influence of the covariate on the response variable is determined by testing the hypothesis that the regression lines for the treatments have a slope of zero. This hypothesis is

$$H_0\text{: } \beta_1 = 0 \quad \text{versus} \quad H_a\text{: } \beta_1 \neq 0$$

for the full model. Our test statistic is based on the sum of squares reduction due to the addition of the covariate x to the model and is given as

$$\text{SS}_{\text{Cov}} = \text{SSE}_{\text{RI}} - \text{SSE}_{\text{F}}$$

We then form the F test

$$F = \frac{\text{SS}_{\text{Cov}}}{\text{SSE}_{\text{F}}/(N - t - 1)}$$

where N is the number of observations in the experiment. Our decision rule is then given by

$$\text{Reject } H_0\text{: } \beta_1 = 0 \quad \text{if} \quad F \geq F_{\alpha,\, 1,\, N-t-1}$$

If we determined that the covariate does have a significant linear relationship with the response variable, we would next test for a significant treatment effect using the adjusted treatment means. That is, we want to test the hypotheses

$$H_0\text{: } \tau_1 = \tau_2 = \cdots = \tau_t = 0 \quad \text{versus} \quad H_a\text{: Not all } \tau_i\text{s are 0.}$$

In the regression model, this is equivalent to testing that the regression lines have the same intercept (β_0). Thus, from Table 16.3, we are testing

$$H_0\text{: } \beta_2 = \beta_3 = \cdots = \beta_t = 0 \quad \text{versus} \quad H_a\text{: Not all of } \beta_2, \beta_3, \ldots, \beta_t \text{ are 0.}$$

Our test statistic is based on the sum of squares reduction due to the addition of the differences in the treatment means to the model and is given

$$\text{SS}_{\text{Trt}} = \text{SSE}_{\text{RII}} - \text{SSE}_{\text{F}}$$

We then form the F test

$$F = \frac{\text{SS}_{\text{Trt}}/(t - 1)}{\text{SSE}_{\text{F}}/(N - t - 1)}$$

Our decision rule is then given by

$$\text{Reject } H_0\text{: } \beta_2 = \beta_3 = \cdots \beta_t = 0 \text{ if } F \geq F_{\alpha,\, t-1,\, N-t-1}$$

If we reject H_0, then we can evaluate treatment differences by examining the estimated adjusted treatment means using the formula

$$\hat{\mu}_{\text{Adj},i} = \bar{y}_{i.} - \hat{\beta}_1(\bar{x}_{i.} - \bar{x}_{..})$$

which adjusts the observed treatment means for the effect of the covariate. This effect is estimated by considering how large a difference exists between the mean value of the covariate observed for the experimental units receiving treatment i and the average value on the covariate over all treatments.

We can also estimate the adjusted treatment means using the regression model. From Table 16.3, for treatments $i = 2, 3, \ldots, t$,

$$\mu_i = E(y) = \beta_0 + \beta_i + \beta_1 x_1$$

and for $i = 1$,

$$\mu_1 = E(y) = \beta_0 + \beta_1 x_1$$

The estimated adjusted treatment means are obtained by estimating the mean value of y for each treatment group corresponding to the overall mean value of the covariate, $x_1 = \bar{x}_{..}$. It follows that

$$\hat{\mu}_{\text{Adj},i} = \hat{\beta}_0 + \hat{\beta}_i + \hat{\beta}_1 \bar{x}_{..}$$

for treatments $i = 2, 3, \ldots, t$ and

$$\hat{\mu}_{\text{Adj},1} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_{..}$$

for treatment 1. The estimated standard error of the estimated ith treatment mean, $\hat{\mu}_{\text{Adj},i}$, is given by

$$\text{SE}(\hat{\mu}_{\text{Adj},i}) = \sqrt{\text{MSE}_{\text{F}}\left(\frac{1}{n} + \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{E_{xx}}\right)}$$

where $E_{xx} = \sum\sum_{ij}(x_{ij} - \bar{x}_{i.})^2$. The estimated standard error of the difference between two adjusted treatment means, $\hat{\mu}_{\text{Adj},i} - \hat{\mu}_{\text{Adj},h}$, is given by

$$\text{SE}(\hat{\mu}_{\text{Adj},i} - \hat{\mu}_{\text{Adj},h}) = \sqrt{\text{MSE}_{\text{F}}\left(\frac{2}{n} + \frac{(\bar{x}_{i.} - \bar{x}_{h.})^2}{E_{xx}}\right)}$$

where MSE_{F} is the MSE from the full model. These estimated standard errors can now be used to place confidence intervals on the adjusted treatment means and their differences.

The following example will illustrate the ideas of analysis of covariance.

EXAMPLE 16.2

Refer to Example 16.1, where we had three treatments—a control (C), a slow-release fertilizer (S), and a fast-release fertilizer (F)—and we used plant height at the beginning of the study as a covariate. Our response variable was the seed yield of peanut plants, and we had 10 replicates.

a. Write the model for an analysis of covariance.
b. Use the computer output shown here to test whether the covariate provides a significant reduction in experimental error.
c. Give the linear regression equations for the three treatment groups.
d. Compute the observed and adjusted treatment means for the three treatment groups.

e. Does there appear to be a significant difference among the three treatments after adjusting for the covariate?

The computer printout for the analysis is given here.

```
FULL MODEL

General Linear Models Procedure

Dependent Variable: Y    YIELD
                                  Sum of        Mean
Source                    DF     Squares      Square    F Value    Pr > F

Model                      3   214.37595    71.45865    4447.85    0.0001

Error                     26     0.41771     0.01607

Corrected Total           29   214.79367

                                T for H0:    Pr > |T|    Std Error of
INTERCEPT      9.529256364         71.34      0.0001      0.13357349
X1 (COV)       0.055809949         20.41      0.0001      0.00273429
X2 (S)         3.571637117         62.62      0.0001      0.05703267
X3 (F)        -3.144155615        -52.08      0.0001      0.06037390
-------------------------------------------------------------------------
REDUCED MODEL I

General Linear Models Procedure

Dependent Variable: Y    YIELD
                                  Sum of        Mean
Source                    DF     Squares      Square    F Value    Pr > F

Model                      2   207.68267   103.84133     394.28    0.0001

Error                     27     7.11100     0.26337

Corrected Total           29   214.79367

                              T for H0:    Pr > |T|    Std Error of
Parameter       Estimate    Parameter = 0                Estimate

INTERCEPT    12.13000000           74.74      0.0001      0.16228690
X2 (S)        3.70000000           16.12      0.0001      0.22950833
X3 (F)       -2.72000000          -11.85      0.0001      0.22950833
-------------------------------------------------------------------------
REDUCED MODEL II

General Linear Models Procedure

Dependent Variable: Y    YIELD
                                  Sum of        Mean
Source                    DF      Squares      Square    F Value    Pr > F

Model                      1    0.4721494   0.4721494       0.06    0.8057

Error                     28   214.3215172   7.6543399

Corrected Total           29   214.7936667

                              T for H0:    Pr > |T|    Std Error of
Parameter       Estimate    Parameter = 0                Estimate

INTERCEPT    13.14900450            4.64      0.0001      2.83300563
X1 (COV)     -0.01387451           -0.25      0.8057      0.05586395
```

Solution

a. We have a completely randomized design with three treatments, 10 replications per treatment, and a single covariate. The model is thus given by $y_{ij} = \mu_i + \beta_1(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij}$, for $i = 1, 2, 3$ and $j = 1, \ldots, 10$.

The full model using regression notation is

Full model (in which the regression lines have different intercepts but a common slope):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

y = yield
x_1 = plant height
$x_2 = 1$ if treatment is S $\quad x_2 = 0$ otherwise
$x_3 = 1$ if treatment is F $\quad x_3 = 0$ otherwise

The expected values of the response for the three treatments are shown here.

Treatment	Expected Responses
C	$\beta_0 + \beta_1 x_1$
S	$(\beta_0 + \beta_2) + \beta_1 x_1$
F	$(\beta_0 + \beta_3) + \beta_1 x_1$

The corresponding reduced models are

Reduced model I (in which the regression lines have a slope equal to zero; that is, the covariate is unrelated to the response variable):

$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Reduced model II (in which the regression lines have a common intercept, β_0, and common slope, β_1):

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

b. We want to test whether the covariate provides a reduction in the experimental error. That is, we need to test that the common slope (β_1) is zero:

$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_a: \beta_1 \neq 0$$

From the computer output,

$$\text{SSE}_{\text{F}} = .41771 \quad \text{SSE}_{\text{RI}} = 7.11100$$

Thus, we have

$$\text{SS}_{\text{Cov}} = \text{SSE}_{\text{RI}} - \text{SSE}_{\text{F}} = 7.111 - .41771 = 6.69329$$

Our F test is

$$F = \frac{6.69329}{.41771/(30 - 3 - 1)} = 416.62 \text{ and } F_{.05,1,26} = 4.23$$

Because 416.62 is greater than 4.23, we reject H_0 and conclude, with p-value $< .0001$, that the plant height (the covariate) is significantly related to plant seed yield (i.e., the slope β_1 is different from zero).

c. From the output for the full model, we obtain the least-squares estimates:

$$\hat{\beta}_0 = 9.53,\ \hat{\beta}_1 = .0558,\ \hat{\beta}_2 = 3.57,\ \hat{\beta}_3 = -3.14$$

The estimated seed yields, with adjustments for initial plant height for the three treatments, are

$$\text{Control:} \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 = 9.53 + .0558x_1$$

$$\text{Slow release:} \quad \hat{y} = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 x_1 = (9.53 + 3.57) + .0558x_1 = 13.1 + .0558x_1$$

$$\text{Fast release:} \quad \hat{y} = (\hat{\beta}_0 + \hat{\beta}_3) + \hat{\beta}_1 x_1 = (9.53 - 3.14) + .0558x_1 = 6.39 + .0558x_1$$

d. The observed sample means are given in Table 16.4.

TABLE 16.4
Sample means for Example 16.2

	Control	Slow Release	Fast Release	Overall
y	12.13	15.83	9.41	12.457
x	46.60	48.90	54.20	49.900

We can obtain the estimated adjusted means by substituting the overall mean plant height for x_1 in the separate regression equations:

$$\text{Control:} \quad \hat{\mu}_{\text{Adj},1} = 9.53 + .0558(49.90) = 12.31$$

$$\text{Slow release:} \quad \hat{\mu}_{\text{Adj},2} = 13.1 + .0558(49.90) = 15.88$$

$$\text{Fast release:} \quad \hat{\mu}_{\text{Adj},3} = 6.39 + .0558(49.90) = 9.17$$

Alternatively, we could obtain the estimated adjusted means using the formula

$$\hat{\mu}_{\text{Adj},i} = \bar{y}_{i.} - \hat{\beta}_1(\bar{x}_{i.} - \bar{x}_{..})$$

$$\text{Control:} \quad \hat{\mu}_{\text{Adj},1} = 12.13 - .0558(46.60 - 49.90) = 12.31$$

$$\text{Slow release:} \quad \hat{\mu}_{\text{Adj},2} = 15.83 - .0558(48.9 - 49.90) = 15.88$$

$$\text{Fast release:} \quad \hat{\mu}_{\text{Adj},3} = 9.41 - .0558(54.20 - 49.90) = 9.17$$

Because the slow-release fertilizer plants had an average plant height less than the overall average height, the observed average seed yield was adjusted upward from 15.83 to 15.88, whereas the fast-release fertilizer's average seed yield was adjusted downward from 9.41 to 9.17.

e. We can test for a difference in the average seed yields of the three treatments by examining the sum of squares error in reduced model II. We want to test the following hypotheses:

H_0: $\mu_{\text{Adj},1} = \mu_{\text{Adj},2} = \ldots = \mu_{\text{Adj},t}$ versus H_a: Not all $\mu_{\text{Adj},i}$s are equal.

This is equivalent to testing the null hypothesis that the regression lines have a common intercept (β_0); that is, we want to test

$$H_0: \ \beta_2 = \beta_3 = 0 \quad \text{versus} \quad H_a: \ \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$$

From the computer output,

$$\text{SSE}_{\text{F}} = .41771 \qquad \text{SSE}_{\text{RII}} = 214.3215$$

Thus, we have

$$SS_{Trt} = SSE_{RII} - SSE_F = 214.3215 - .41771 = 213.90$$

Our F test thus is

$$F = \frac{213.90/(3-1)}{.41771/(30-3-1)} = 6{,}657.13 \text{ and } F_{.05,2,26} = 3.37$$

Because 6,657.13 is greater than 3.37, we reject H_0 and conclude, with p-value $< .0001$, that the intercepts are not equal, and, hence, there is significant evidence of a difference in the adjusted plant seed yields for the three types of fertilizers. ■

The conclusions we reached in Example 16.2 are dependent on the validity of the conditions we placed on the model. We can evaluate the condition of independent and homogeneous, normally distributed error terms by examining the residuals from the fitted model:

$$e_{ij} = y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_{1ij} - \hat{\beta}_2 x_{2ij} - \cdots - \hat{\beta}_t x_{tij}$$

We can then apply plots and tests of normality to the e_{ij}s to evaluate the equal variance and normality conditions.

The three added conditions for the analysis of covariance are evaluated in the following manner.

The Relationship Between the Response and the Covariate Is Linear We can evaluate this condition as we did in regression analysis through the use of plots and tests of hypotheses. We can plot y versus x separately for each treatment and assess whether the plotted points follow a straight line. A separate regression line can be fitted for each treatment using the methods of Chapter 12. We can then assess the residuals from the t fitted lines and conduct tests of lack of fit to determine whether any of the t fitted lines need higher-order terms in the covariate x_{ij}. The situation of higher-order relationships will be discussed in Section 16.4.

The Regression (Slope) Coefficient Is the Same for All t Treatments Consider the following model:

$$\text{Model A: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_r x_t + \beta_{t+1} x_1 x_2 + \beta_{t+2} x_1 x_3 + \cdots + \beta_{2t-1} x_1 x_t + \varepsilon$$

where $x_2, \ldots, x_t$ are the indicator variables for the treatments and x_1 is the covariate. This regression model yields separate regression lines, with possibly different slopes and different intercepts, for each treatment. (See the expected responses for model A shown in Table 16.5.)

We next consider a reduced model, in which we require the slopes to be the same for all treatments but allow for different intercepts.

$$\text{Model B: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_r x_t + \varepsilon$$

TABLE 16.5 Expected values for model A

Treatment	Expected Response
1	$\mu_1 = \beta_0 + \beta_1 x_1$
2	$\mu_2 = (\beta_0 + \beta_2) + (\beta_1 + \beta_{t+1}) x_1$
3	$\mu_3 = (\beta_0 + \beta_3) + (\beta_1 + \beta_{t+2}) x_1$
⋮	⋮
t	$\mu_t = (\beta_0 + \beta_t) + (\beta_1 + \beta_{2t-1}) x_1$

The test for equal slopes would involve testing

$$H_0: \beta_{t+1} = \beta_{t+2} = \cdots = \beta_{2t-1} = 0$$

$$H_a: \text{At least one of } \beta_{t+1}, \beta_{t+2}, \ldots, \beta_{2t-1} \text{ is not 0.}$$

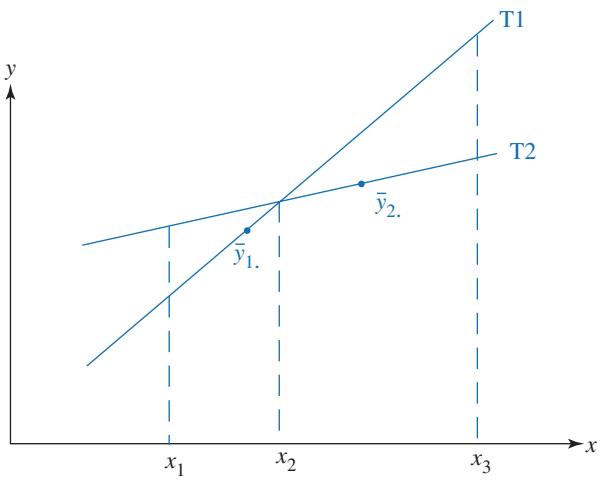
The test statistic would be obtained by fitting models A and B.

$$F = \frac{(\text{SSE}_\text{B} - \text{SSE}_\text{A})/(t-1)}{\text{SSE}_\text{A}/(N-2t)} \quad \text{with } \text{df}_1 = t-1, \text{df}_2 = N-2t$$

This would determine whether the regression lines relating the response to the covariate have the same slope. This is a crucial assumption because if the slopes are different, then the difference in the adjusted treatment means is highly dependent on the level of the covariate chosen for adjustment. This situation is similar to experiments in which we have two factors with significant interactions and inferences about one factor depending on the level of the second factor. The situation in which the lines relating the response to the covariate have different slopes is displayed in Figure 16.3. From this figure, we can observe that amount of adjustment varies greatly depending on which treatment and which value of the covariate are selected for adjustment.

When the treatments have different slopes, then our conclusion concerning which treatment has the largest (smallest) adjusted treatment mean depends on the value of the covariate. In Figure 16.3, when the covariate has value x_1, treatment $T2$ has a larger estimated mean response; at x_2, the two estimated mean responses are equal; and at x_3, treatment $T1$ has a larger mean response than does treatment $T2$. This situation is considerably different from the case in which the treatments have the same slope. With equal slopes, the difference between the treatments remains consistent across the values of the covariate. When the treatments have different slopes, then the differences between the treatments vary depending on the value of the covariate. Thus, all conclusions about the difference in the treatments must be made conditional on the value of the covariate. In this situation, the researcher provides a value of the covariate; then comparisons of the adjusted treatment means can be made. This process is repeated over as many values of the covariate as are of interest to the researcher. Of course, multiple comparison adjustments to the type I error rates must be made.

FIGURE 16.3
Regression lines relating the response and covariate with different slopes



The Treatments Do Not Affect the Covariate, x_{ij} In experiments where both the covariate x and the response variable y are affected by the treatments, we cannot validly apply the methods of analysis of covariance. The appropriate method of analysis would involve multivariate analysis where we treat the response as a bivariate variable (x, y). When the covariate is measured prior to the random assignment of treatments to the experimental units, the analysis of covariance model would be appropriate because it would be impossible for the treatment to affect the covariate. When the covariate is measuring conditions in the experimental setting—that is, the covariate is measured during the running of the experiment—the experimenter must decide whether the treatments have an affect on the covariate. Only after the experimenter determines that the treatments have not affected the covariate can we correctly adjust the treatment means for the covariate.

EXAMPLE 16.3

Refer to Example 16.1. Evaluate the necessary conditions in the analysis of covariance model, using the computer output given here.

```
MODEL A: DIFFERENT SLOPES FOR EACH TREATMENT

General Linear Models Procedure
Number of observations in data set = 30

Dependent Variable: Y    YIELD
                                  Sum of          Mean
Source                     DF       Squares        Square    F Value    Pr > F
Model                       5     214.43722      42.88744    2887.70    0.0001
Error                      24       0.35644       0.01485
Corrected Total            29     214.79367

Source                     DF   Type III SS   Mean Square   F Value   Pr > F

X1                          1     2.6167178     2.6167178    176.19    0.0001
X2                          1     2.5905994     2.5905994    174.43    0.0001
X3                          1     1.4990044     1.4990044    100.93    0.0001
X2*X1                       1     0.0190292     0.0190292      1.28    0.2688
X3*X1                       1     0.0151538     0.0151538      1.02    0.3225

                                 T for H0:    Pr > |T|    Std Error of
Parameter        Estimate    Parameter = 0                  Estimate
INTERCEPT     9.491768741           46.88      0.0001      0.20245904
X1            0.056614405           13.27      0.0001      0.00426518
X2            3.906558043           13.21      0.0001      0.29578964
X3           -3.519620102          -10.05      0.0001      0.35033468
X2*X1        -0.006886936           -1.13      0.2688      0.00608421
X3*X1         0.006814587            1.01      0.3225      0.00674632
-------------------------------------------------------------------------------
MODEL B: SAME SLOPE FOR ALL TREATMENTS

General Linear Models Procedure
Number of observations in data set = 30

Dependent Variable: Y    YIELD
                                  Sum of          Mean
Source                     DF       Squares        Square    F Value    Pr > F
Model                       3     214.37595      71.45865    4447.85    0.0001
Error                      26       0.41771       0.01607
Corrected Total            29     214.79367
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	6.693287	6.693287	416.62	0.0001
X2	1	63.007424	63.007424	3921.82	0.0001
X3	1	43.572654	43.572654	2712.12	0.0001

Parameter	Estimate	T for H0: Parameter = 0	Pr > |T|	Std Error of Estimate
INTERCEPT	9.529256364	71.34	0.0001	0.13357349
X1	0.055809949	20.41	0.0001	0.00273429
X2	3.571637117	62.62	0.0001	0.05703267
X3	-3.144155615	-52.08	0.0001	0.06037390

Solution From Figure 16.2, we can see that the lines relating seed yield to plant height for the three treatments appear to be adequately fit by a straight line and the three slopes appear to be the same; that is, we have three parallel lines with possibly different intercepts. The computer output is obtained by fitting model A (different slopes and different intercepts) and model B (same slopes but different intercepts) to the plant seed yield data.

From the output, we can compute

$$F = \frac{(\text{SSE}_\text{B} - \text{SSE}_\text{A})/(t-1)}{\text{SSE}_\text{A}/(N-2t)} = \frac{(.41771 - .35644)/(3-1)}{.35644/(30-6)} = 2.06$$

with $\text{df}_1 = 2$ and $\text{df}_2 = 24$. Because $F_{.05,2,24} = 3.40$, we fail to reject H_0 and conclude, with p-value = .1494, that there is not significant evidence of a difference in the slopes of the three lines. Because the covariate, plant height, was measured prior to assigning the type of fertilizer to the plants, the treatments cannot have an affect on the covariate. The remaining conditions of equal variance and normality can be assessed using a residual analysis. ■

16.3 The Extrapolation Problem

In the previous section, we discussed how to compare two (or more) treatments from a completely randomized design with one covariable. If the regression equations for the treatments are linear in terms of the covariable and parallel, we said we could compare the treatments using the adjusted treatment means. However, as with most methods, the analysis of covariance methods should not be used blindly. Even if the linearity and parallelism assumptions hold, we can have problems if the values of the covariable do not have considerable overlap for the treatment groups. We will illustrate this with an example.

Suppose that we were interested in comparing self-esteem scores for alcoholics and drug addicts. We collected a sample of nine alcoholics and a sample of nine drug addicts, and for each individual, we obtained his or her self-esteem score and age. The data are shown in Table 16.6.

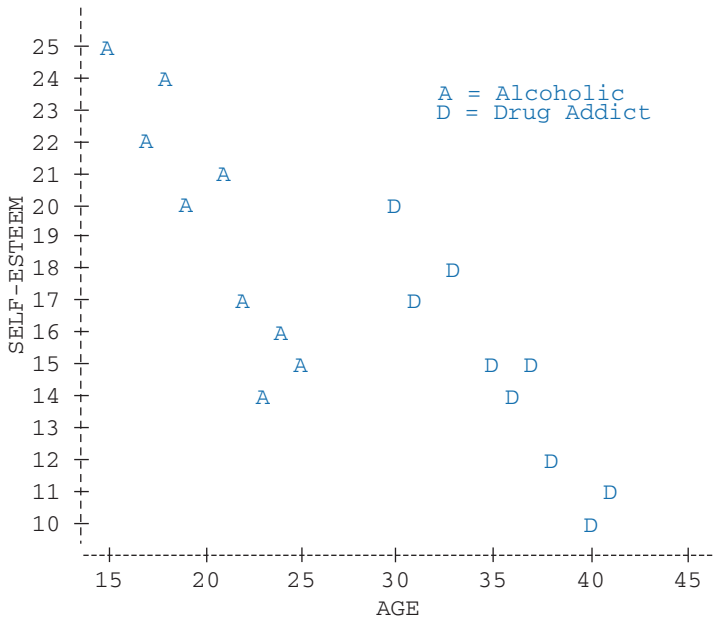
If we blindly followed the analysis of covariance procedures without looking at the data, we would find the regression equations for alcoholics and drug addicts to be reasonably linear and parallel. From the computer output displayed in Figure 16.4, we would note from the plotted data that the data values for alcoholics (A) would fall near a straight line, as would the points for drug addicts (D). If we used the sum of squares error for the two models, we would obtain

$$F = \frac{(30.88 - 27.39)/(2-1)}{1.9567} = 1.78$$

TABLE 16.6
Self-esteem scores and ages for a sample of alcoholics and drug addicts

Alcoholics		Drug Addicts	
Self-Esteem	**Age**	**Self-Esteem**	**Age**
25	15	20	30
22	17	17	31
24	18	18	33
20	19	15	35
21	21	14	36
17	22	15	37
14	23	12	38
16	24	10	40
15	25	11	41

FIGURE 16.4
Self-esteem scores as a function of age



with $df_1 = 1$ and $df_2 = 14$. The p-value for the observed F-value would be $P(F \geq 1.78) = 0.2035$. Thus, we would fail to reject the hypothesis that the slopes of the lines relating self-esteem to age are the same for the alcoholics and the drug addicts. Furthermore, from the computer output for model B, we would find that the p-value for testing a difference in the adjusted mean self-esteem scores is $P(F \geq 34.14) < 0.0001$. The two groups of addicts would appear to have different adjusted mean self-esteem scores.

```
MODEL A: DIFFERENT SLOPES AND TREATMENT DIFFERENCES

Dependent Variable: Y    SELF-ESTEEM

                                  Sum of          Mean
Source                    DF      Squares        Square    F Value    Pr > F
Model                      3    286.60611      95.53537      48.82    0.0001
Error                     14     27.39389       1.95671
Corrected Total           17    314.00000

Source                    DF   Type III SS   Mean Square   F Value    Pr > F
X1                         1    188.51593     188.51593      96.34    0.0001
X2                         1      0.43265       0.43265       0.22    0.6454
X2*X1                      1      3.48284       3.48284       1.78    0.2035
```

```
                                T for H0:    Pr > |T|    Std Error of
Parameter       Estimate     Parameter = 0                   Estimate

INTERCEPT    44.18390805 B            9.49      0.0001     4.65570471
X1           -0.82758621 B           -6.37      0.0001     0.12987748
X2           -2.60800443 B           -0.47      0.6454     5.54628759
X2*X1        -0.26036560 B           -1.33      0.2035     0.19515497
------------------------------------------------------------------------
MODEL B: SAME SLOPES AND TREATMENT DIFFERENCES

Dependent Variable: Y    SELF-ESTEEM

                                  Sum of          Mean
Source                   DF       Squares        Square    F Value    Pr > F
Model                     2     283.12327     141.56163      68.77    0.0001
Error                    15      30.87673       2.05845
Corrected Total          17     314.00000

Source                   DF    Type III SS    Mean Square    F Value    Pr > F
X1                        1     185.12327     185.12327      89.93    0.0001
X2                        1      70.27928      70.27928      34.14    0.0001

                                T for H0:    Pr > |T|    Std Error of
Parameter       Estimate     Parameter = 0                   Estimate

INTERCEPT    48.29686944 B           13.50      0.0001     3.57834982
X1           -0.94290288             -9.48      0.0001     0.09942750
X2           -9.68641053 B           -5.84      0.0001     1.65775088

REDUCED MODEL I: TREATMENT DIFFERENCES WITH NO COVARIATE

General Linear Models Procedure

Dependent Variable: Y    SELF-ESTEEM
                                  Sum of          Mean
Source                   DF       Squares        Square    F Value    Pr > F
Model                     1      98.000000      98.000000       7.26    0.0160
Error                    16     216.000000      13.500000
Corrected Total          17     314.000000

Source                   DF    Type III SS    Mean Square    F Value    Pr > F
X2                        1      98.000000      98.000000       7.26    0.0160
------------------------------------------------------------------------
REDUCED MODEL II: COVARIATE BUT NO TREATMENT DIFFERENCES

General Linear Models Procedure

Dependent Variable: Y    SELF-ESTEEM
                                  Sum of          Mean
Source                   DF       Squares        Square    F Value    Pr > F
Model                     1     212.84398      212.84398      33.67    0.0001
Error                    16     101.15602       6.32225
Corrected Total          17     314.00000


Source                   DF    Type III SS    Mean Square    F Value    Pr > F
X1                        1     212.84398      212.84398      33.67    0.0001

                              T for H0:    Pr > |T|    Std Error of
Parameter       Estimate   Parameter = 0                   Estimate
INTERCEPT    28.57258960           13.73      0.0001     2.08069635
X1           -0.41248834           -5.80      0.0001     0.07109137
```

Do alcoholics and drug addicts really have different self-esteem scores? One possible explanation for the difference in scores is that we are dealing with two different age groups: The alcoholics sampled ranged in age from 15 to 25 years, whereas the drug addicts were between the ages of 30 and 41. This difference in ages for the two groups is borne out in the scatterplot shown in Figure 16.4.

The mean ages for the alcoholics and drug addicts are 20.44 and 35.67 years, respectively, while the combined mean age is 28.06 years. Note that the combined mean is outside the age range for each of the separate samples. We have no information about self-esteem scores for drug addicts under 30 years of age and no information about self-esteem scores for alcoholics above the age of 25. Hence, it would be inappropriate to compare the predicted self-esteem scores at the "adjusted" age (28.06) because this involves an extrapolation beyond the ages observed for the separate samples. For this example, it would be difficult to make any comparison between the alcoholics and drug addicts because of the age differences and other possible (unmeasured) differences between the two groups.

In situations where there is the potential for the ranges of values for the covariate to not have a considerable overlap, how should a researcher design the study to avoid the problems described above? When designing the study, examine the value of the covariate for each experimental unit, and if the range of values is large, then use a randomized block design to assign the experimental units to the treatments. In the above study, the researcher could have avoided the confounding of age group with type of addiction by blocking on age prior to measuring the self-esteem of the participants. This design would consist of two stratified random samples, one from the population of people who were alcoholics and the other from the population of drug addicts. The stratification would be based on age—with three or four age groups. This would then guarantee that there would be considerable overlap of the ages over the two types of addiction. We will discuss how to analyze an experiment in which both blocking and covariates are present in the next section.

So don't forget to look at your data. The potential for extrapolation, although not as obvious as for our example, should become apparent with plots of the data. Then you can avoid using an analysis of covariance to make comparisons of adjusted treatment means (or, in fact, any comparison) when the adjustment may be inappropriate. These same problems can occur with the extensions of these methods to include more than one covariable and more complicated experimental designs—but it is more difficult to detect the problem.

16.4 Multiple Covariates and More Complicated Designs

The sample procedures discussed in Section 16.2 can also be applied to completely randomized designs with one or more covariates. Including more than one covariate in the model merely means that we have more than one quantitative independent variable in our model. For example, we might wish to compare the social status y of several different occupational groups while incorporating information on the number of years x_1 of formal education beyond high school and the income level x_2 of each individual in a group. As mentioned previously, we need not restrict ourselves to linear terms in the covariate(s). Thus, we might have a response related to two covariates (x_1 and x_2) and $t = 3$ treatments using the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \beta_8 x_1^2 x_3 + \beta_9 x_1^2 x_4 + \beta_{10} x_2 x_3 + \beta_{11} x_2 x_4 + \varepsilon$$

where

$x_3 = 1$ if treatment 2 $\quad x_3 = 0$ otherwise
$x_4 = 1$ if treatment 3 $\quad x_4 = 0$ otherwise

We can readily obtain an interpretation of the βs by using a table of expected values similar to Table 16.3.

EXAMPLE 16.4

For the model with $t = 3$ treatments, two covariates, (x_1 and x_2) and the response equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \beta_8 x_1^2 x_3 + \beta_9 x_1^2 x_4 + \beta_{10} x_2 x_3 + \beta_{11} x_2 x_4 + \varepsilon$$

relate the parameters in the model to the expected responses for each of the treatments.

Solution The table of expected values is given in Table 16.7.

TABLE 16.7
Expected responses for the model in Example 16.4

Treatment	Expected Response
1	$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$
2	$(\beta_0 + \beta_4) + (\beta_1 + \beta_6)x_1 + (\beta_2 + \beta_8)x_1^2 + (\beta_3 + \beta_{10})x_2$
3	$(\beta_0 + \beta_5) + (\beta_1 + \beta_7)x_1 + (\beta_2 + \beta_9)x_1^2 + (\beta_3 + \beta_{11})x_2$

Thus, the y-intercepts of the three adjusted treatment lines for treatments 1, 2, and 3 are β_0, $\beta_0 + \beta_4$, and $\beta_0 + \beta_5$, respectively. Similarly, the partial slopes for the covariate x_1 are β_1, $\beta_1 + \beta_6$, and $\beta_1 + \beta_7$, respectively. The partial slopes for the covariate x_1^2 are β_2, $\beta_2 + \beta_8$, and $\beta_2 + \beta_9$, respectively. The partial slopes for the covariate x_2 are β_3, $\beta_3 + \beta_{10}$, and $\beta_3 + \beta_{11}$, respectively. The hypotheses for testing for differences in the partial slopes for x_1 would be

$$H_0: \beta_6 = 0, \beta_7 = 0 \quad \text{versus} \quad H_a: \beta_6 \neq 0 \text{ and/or } \beta_7 \neq 0$$

The hypotheses for testing for differences in the partial slopes for x_1^2 would be

$$H_0: \beta_8 = 0, \beta_9 = 0 \quad \text{versus} \quad H_a: \beta_8 \neq 0 \text{ and/or } \beta_9 \neq 0$$

The hypotheses for testing for differences in the partial slopes for x_2 would be

$$H_0: \beta_{10} = 0, \beta_{11} = 0 \quad \text{versus} \quad H_a: \beta_{10} \neq 0 \text{ and/or } \beta_{11} \neq 0$$

If one or more of the three null hypotheses are rejected, then we would conclude that the adjusted treatment mean planes are not parallel and conclusions about treatment differences cannot be made without specifying values of the covariates. ■

An analysis of covariance for more-complicated designs can also be obtained using general linear model methodology. The techniques for handling adjustments for covariates in randomized complete block designs and Latin squares are similar to the methods we discussed for completely randomized designs. The following example will illustrate the modeling for a randomized complete block design.

EXAMPLE 16.5

Suppose we have a randomized complete block design with two blocks, three treatments, one covariate x, and $n > 1$ observations per treatment in each block. Write the model for this experimental situation, assuming the response is linearly related to the covariate for each treatment. Identify the parameters in the model.

Solution The model is written as

$$y_{ijk} = \beta_0 + \gamma_i + \tau_j + \delta_j x_{ijk} + \varepsilon_{ijk}$$

where $i = 1, 2, 3; j = 1, 2$; and $k = 1, \ldots, n$. The parameters are identified as follows: β_0 is the intercept of the regression of y on x, τ_j is the jth treatment effect, γ_i is the

ith block effect, δ_j is the slope of the regression of y on x for treatment j, and the ε_{ijk}s are the random error variables. We can write this in a generalized linear model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_3 + \beta_6 x_1 x_4 + \varepsilon$$

where

x_1 = covariate
$x_2 = 1$ if block 2 $\quad x_2 = 0$ otherwise
$x_3 = 1$ if treatment 2 $\quad x_3 = 0$ otherwise
$x_4 = 1$ if treatment 3 $\quad x_4 = 0$ otherwise

We immediately recognize this as a model relating a response y to a quantitative variable x_1 and two qualitative variables: blocks and treatments. An interpretation of the βs in the model is obtained from the table of expected responses shown in Table 16.8.

TABLE 16.8
Expected values for the randomized block design with one covariate

Block	Treatment	Expected Response
1	1	$\beta_0 + \beta_1 x_1$
	2	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
	3	$(\beta_0 + \beta_4) + (\beta_1 + \beta_6)x_1$
2	1	$(\beta_0 + \beta_2) + \beta_1 x_1$
	2	$(\beta_0 + \beta_2 + \beta_3) + (\beta_1 + \beta_5)x_1$
	3	$(\beta_0 + \beta_2 + \beta_4) + (\beta_1 + \beta_6)x_1$

The model we formulated in Example 16.5 not only provides for a linear relationship between y and x_1 for each of the treatments in each block but also allows for differences among intercepts and slopes. If we wanted to test for the equality of the slopes across treatments, we would use the null hypothesis

$$H_0: \quad \beta_5 = \beta_6 = 0$$

If there is insufficient evidence to reject H_0, we would proceed with the reduced model (obtained by setting $\beta_5 = \beta_6 = 0$ in our model)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

A test for differences among treatments adjusted for the covariate, when slopes are equal, could be obtained by fitting a complete and a reduced model for the null hypothesis

$$H_0: \quad \beta_3 = \beta_4 = 0$$

16.5 RESEARCH STUDY: Evaluation of Cool-Season Grasses for Putting Greens

The objective of the study was to compare the mean speed of putted golf balls on three cultivars used on golf course greens. In Section 16.1 we described the research problem and why the study was being conducted. The next step in the process would be designing the data collection process.

Designing the Data Collection

The researchers considered the following issues in designing an appropriate experiment to evaluate the cultivars:

1. What performance measures should be used to evaluate the cultivars?
2. Does the geographical region of the country affect the performance of the cultivar?

3. Do the cultivars perform differently during differing times of the golf season?
4. What soil factors affect the performance characteristics of the cultivars?
5. How many replications per cultivar are needed to obtain a reliable estimate of cultivar performance?
6. What environmental factors may affect the performance of the cultivars during the test period?
7. What are the valid statistical procedures for evaluating differences in the cultivars?
8. What type of information should be included in a final report to document the differences in the suitability of the cultivars for use on golf course putting greens?

The experiment was conducted, and the data were given in Table 16.1. A plot of the data was presented in Figure 16.1.

Analyzing the Data

From the plot in Figure 16.1, it would appear that the response variable, speed of putted ball, was linearly related to relative humidity, with similar slope coefficients for the three cultivars. We will model the data, evaluate the model conditions, and then test for differences in the adjusted mean speeds for the three cultivars. Because there were regional differences in soil characteristics and climatic conditions, eight different regions of the country were selected for testing sites. At each site, there was a single green for each of the three cultivars. A covariate, relative humidity, was recorded during the time when the speed measurements were obtained on each green. Thus, we have a randomized complete block design with eight blocks (region of country), three treatments (cultivars), and a single covariate (relative humidity). We'll assume a model that relates the response variable (speed of green) to the blocks, treatments, and covariate and that allows for different slopes for the treatments (cultivars) within a region; however, we'll also assume that a green treatment has the same slope across regions.

Model I: Region and cultivar differences with covariate having different slopes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_9 x_1 + \beta_{12} x_{10} x_1 + \varepsilon$$

where

x_1 = relative humidity (covariate)

$x_2 = 1$ if region 1 is used	$x_2 = 0$ otherwise
$x_3 = 1$ if region 2 is used	$x_3 = 0$ otherwise
$x_4 = 1$ if region 3 is used	$x_4 = 0$ otherwise
$x_5 = 1$ if region 4 is used	$x_5 = 0$ otherwise
$x_6 = 1$ if region 5 is used	$x_6 = 0$ otherwise
$x_7 = 1$ if region 6 is used	$x_7 = 0$ otherwise
$x_8 = 1$ if region 7 is used	$x_8 = 0$ otherwise
$x_9 = 1$ if cultivar 1 is used	$x_9 = 0$ otherwise
$x_{10} = 1$ if cultivar 2 is used	$x_{10} = 0$ otherwise

The expected values for model I are shown in Table 16.9.

TABLE 16.9
Expected values for model I in the case study

	Cultivar		
Region	**1**	**2**	**3**
1	$(\beta_0 + \beta_2 + \beta_9) + (\beta_1 + \beta_{11})x_1$	$(\beta_0 + \beta_2 + \beta_{10}) + (\beta_1 + \beta_{12})x_1$	$(\beta_0 + \beta_2) + \beta_1 x_1$
2	$(\beta_0 + \beta_3 + \beta_9) + (\beta_1 + \beta_{11})x_1$	$(\beta_0 + \beta_3 + \beta_{10}) + (\beta_1 + \beta_{12})x_1$	$(\beta_0 + \beta_3) + \beta_1 x_1$
⋮	⋮	⋮	⋮
7	$(\beta_0 + \beta_8 + \beta_9) + (\beta_1 + \beta_{11})x_1$	$(\beta_0 + \beta_8 + \beta_{10}) + (\beta_1 + \beta_{12})x_1$	$(\beta_0 + \beta_8) + \beta_1 x_1$
8	$(\beta_0 + \beta_9) + (\beta_1 + \beta_{11})x_1$	$(\beta_0 + \beta_{10}) + (\beta_1 + \beta_{12})x_1$	$\beta_0 + \beta_1 x_1$

Note that the cultivars have different slopes but that each cultivar has the same slope across regions.

To test whether the linear relationship between speed of putted ball and relative humidity is the same for the three cultivars—that is, whether the three lines have equal slopes—we fit a model to the data in which the three lines have the same slope but different intercepts.

Model II: Region and cultivar differences with covariate having equal slopes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \varepsilon$$

The computer output from fitting these two models is given here.

```
MODEL I: REGION AND TREATMENT DIFFERENCES WITH COVARIATE HAVING UNEQUAL SLOPES

                              The GLM Procedure

Dependent Variable: S    SPEED
                                    Sum of          Mean
Source                     DF       Squares        Square    F Value    Pr > F
Model                      12    18.57446432    1.54787203      54.57    <.0001
Error                      11     0.31203151    0.02836650
Corrected Total            23    18.88649583

Source                     DF    Type III SS    Mean Square    F Value    Pr > F
X1                          1     0.84623766     0.84623766      29.83    0.0002
X2                          1     0.21498101     0.21498101       7.58    0.0188
X3                          1     0.18539490     0.18539490       6.54    0.0267
X4                          1     0.13629629     0.13629629       4.80    0.0508
X5                          1     0.27240763     0.27240763       9.60    0.0101
X6                          1     0.05024586     0.05024586       1.77    0.2101
X7                          1     0.00154873     0.00154873       0.05    0.8195
X8                          1     0.01972964     0.01972964       0.70    0.4220
X9                          1     0.48434458     0.48434458      17.07    0.0017
X10                         1     0.02495287     0.02495287       0.88    0.3684
X1*X9                       1     0.09110959     0.09110959       3.21    0.1006
X1*X10                      1     0.13467594     0.13467594       4.75    0.0520
------------------------------------------------------------------------------
MODEL II: REGION AND TREATMENT DIFFERENCES WITH COVARIATE HAVING EQUAL SLOPES

                              The GLM Procedure

Dependent Variable: S    SPEED
                                    Sum of          Mean
Source                     DF       Squares        Square    F Value    Pr > F
Model                      10    18.41522972    1.84152297      50.80    <.0001
Error                      13     0.47126611    0.03625124
Corrected Total            23    18.88649583
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	3.12245889	3.12245889	86.13	<.0001
X2	1	0.09497230	0.09497230	2.62	0.1295
X3	1	0.15887210	0.15887210	4.38	0.0565
X4	1	0.16679129	0.16679129	4.60	0.0514
X5	1	0.23014999	0.23014999	6.35	0.0256
X6	1	0.04177024	0.04177024	1.15	0.3026
X7	1	0.01075404	0.01075404	0.30	0.5952
X8	1	0.04107013	0.04107013	1.13	0.3065
X9	1	14.08930137	14.08930137	388.66	<.0001
X10	1	3.74477234	3.74477234	103.30	<.0001

A test for equal slopes is obtained by testing in model I the hypotheses

$$H_0: \beta_{11} = \beta_{12} = 0 \quad \text{versus} \quad H_a: \beta_{11} \neq 0 \text{ and/or } \beta_{12} \neq 0$$

The test statistic for H_0 versus H_a is

$$F = \frac{(\text{SSE}_{\text{II}} - \text{SSE}_{\text{I}})/(\text{df}_{\text{EII}} - \text{df}_{\text{EI}})}{\text{MSE}_{\text{I}}} = \frac{(.4713 - .3120)/(13 - 11)}{.0284} = 2.80$$

The p-value is given by $P(F_{2,11} \geq 2.80) = .1040$. Thus, the data support the hypothesis that the three cultivars have the same slope. Next, we can test for differences in the adjusted means of the three cultivars. We fit a model in which the covariate has equal slopes for the three cultivars, but we remove any differences in the cultivars and retain differences due to the blocking variable, regions.

Model III: Covariate with equal slopes, region differences, but no cultivar differences

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \varepsilon$$

The computer output from fitting this model is given here.

MODEL III: COVARIATE WITH EQUAL SLOPES, REGION DIFFERENCES, BUT NO TREATMENT DIFFERENCES

The GLM Procedure

Dependent Variable: S SPEED

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	4.32038575	0.54004822	0.56	0.7971
Error	15	14.56611008	0.97107401		
Corrected Total	23	18.88649583			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	2.00762325	2.00762325	2.07	0.1710
X2	1	0.11376059	0.11376059	0.12	0.7369
X3	1	0.15977479	0.15977479	0.16	0.6907
X4	1	0.21374424	0.21374424	0.22	0.6457
X5	1	0.41270875	0.41270875	0.43	0.5243
X6	1	0.00400793	0.00400793	0.00	0.9496
X7	1	0.00004833	0.00004833	0.00	0.9945
X8	1	0.00350604	0.00350604	0.00	0.9529

A test for differences in the adjusted cultivar means is a test of

$$H_0: \mu_{\text{Adj, C1}} = \mu_{\text{Adj, C2}} = \mu_{\text{Adj, C3}} \quad \text{versus} \quad H_a: \mu_{\text{Adj, C}} \textit{ is} \text{ not all equal.}$$

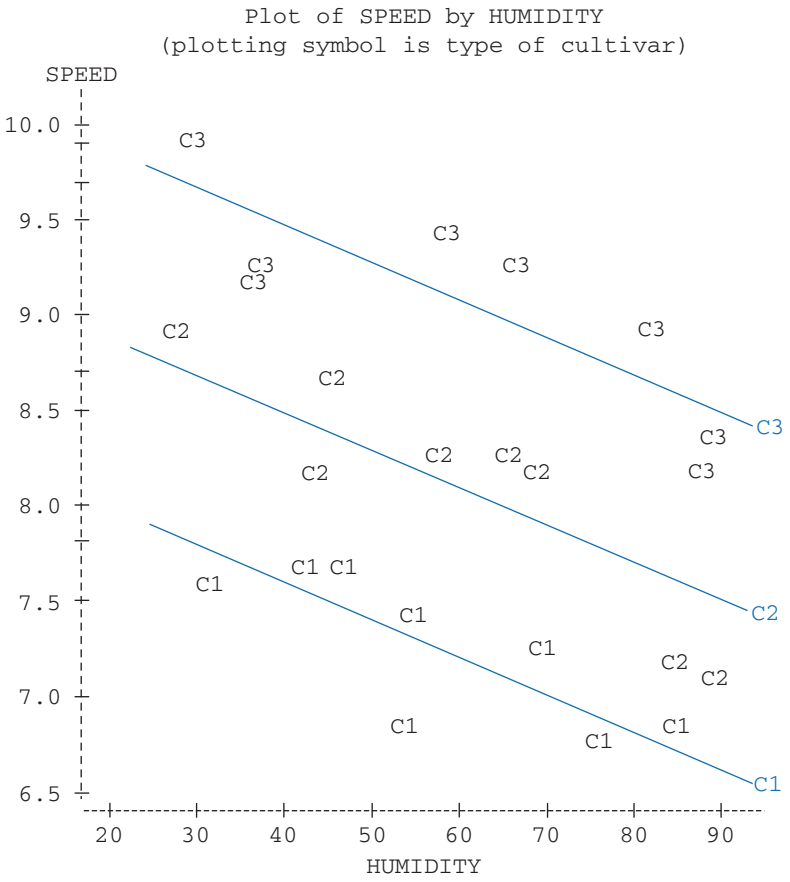
This set of hypotheses is equivalent to testing in model II the hypotheses

$$H_0: \beta_9 = \beta_{10} = 0 \quad \text{versus} \quad H_a: \beta_9 \neq 0 \text{ and/or } \beta_{10} \neq 0$$

The test statistic for H_0 versus H_a is

$$F = \frac{(\text{SSE}_{\text{III}} - \text{SSE}_{\text{II}})/(\text{df}_{\text{EIII}} - \text{df}_{\text{EII}})}{\text{MSE}_{\text{II}}} = \frac{(14.5661 - .4713)/(15 - 13)}{.0363} = 194.14$$

FIGURE 16.5
Cultivar speeds plotted versus relative humidity readings along with fitted lines from the regression model



The p-value is given by $P(F_{2,13} \geq 194.14) < .0001$. Thus, the data strongly support the research hypothesis that there is a difference in the adjusted mean speeds for the three cultivars. We can further investigate what type of differences exist in the three cultivars by examining the plot of the speed and relative humidity data values in Figure 16.5. The lines drawn through the data values were obtained from the parameter estimates in model II. We can observe that cultivar C3 consistently yields higher speeds than the other two cultivars, with cultivar C2 yielding higher speeds than cultivar C1.

The estimated adjusted mean speeds are given in Table 16.10 along with their estimated standard errors, which were used to construct 95% confidence intervals on the mean speeds. From the results in Table 16.10, cultivar C3 has an adjusted mean speed about one unit larger than that of cultivar C2, which has an adjusted mean speed about one unit larger than that of cultivar C1. Differences of this size in the mean speed are considered to be practical differences and will greatly assist golf course designers in selecting the proper cultivar for their course.

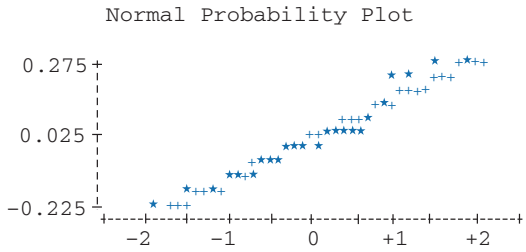
TABLE 16.10
Estimated adjusted cultivar speeds with 95% confidence intervals

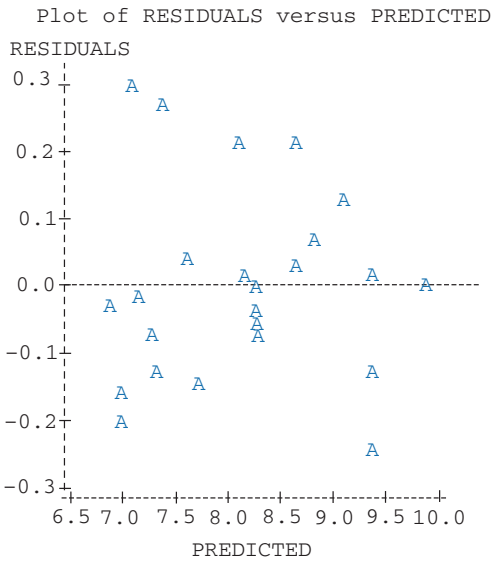
Cultivar	$\hat{\mu}_{\text{Adj}}$	SE($\hat{\mu}_{\text{Adj}}$)	95% Confidence Interval
C1	7.20	.0676	(7.05, 7.35)
C2	8.13	.0674	(7.98, 8.28)
C3	9.08	.0674	(8.93, 9.23)

Prior to using the results obtained above, the researchers must check whether the conditions placed on the analysis of covariance model are satisfied in this experiment. An examination of the following plots of the residuals and plots of the observed data will assist in checking on the validity of the model conditions. The computer printouts of the plots and analysis of the residuals from model II are given here.

```
Univariate Procedure
Variable-RESIDUALS
                  Moments
N                      24   Sum Wgts          24
Mean                    0   Sum                0
Std Dev          0.142759   Variance     0.02038
Skewness         0.522974   Kurtosis   -0.22996

W:Normal         0.954191   Pr<W          0.3405
```

```
Variable-RESIDUALS
     Stem Leaf                     #     Boxplot
        2  79                     2
        2  12                     2
        1                         1
        1  3                      1
        0  8                      1
        0  11134                  5
       -0  220                    3
       -0  8755                   4
       -1  433                    3
       -1  6                      1
       -2  40                     2
          ----+----+----+----+
     Multiply Stem.Leaf by 10**-1
```





The boxplot and stem-and-leaf plot of the residuals do not indicate any extreme values. The normal probability plot indicates that a few residuals are somewhat deviant from the fitted line. However, the test of normality yields a p-value of .3405, so there is strong support for the normality of the residuals. The plot of the residuals versus predicted values does not indicate a violation of the equal variances of the residuals assumption because the spread in the residuals remains reasonably constant across the predicted values. The equal slopes assumption was tested and found to be satisfied. From the plotted values in Figure 16.5, we can observe that there is a linear relationship between speed and relative humidity. Thus, it would appear that the requisite conditions for an analysis of covariance have not been violated in this experiment.

16.6 Summary

In this chapter, we presented a procedure called the analysis of covariance. Here, for each value of y, we also observe a value of concomitant variable x. This second variable, called a covariate, is recognized as an uncontrolled quantitative independent variable. Because of this fact, we can formulate models using the general linear model methodology of previous chapters.

In most situations when reference is made to an analysis of covariance, it is assumed that the response is linearly related to the covariate x, with the slope of the line the same for all treatment groups. Then a test for treatments adjusted for the covariate is performed. Actually, many people run analyses of covariance without checking the assumptions of parallelism. Rather than trying to force a particular model onto an experimental situation, it would be much better to postulate a reasonable (not necessarily linear) model relating the response y to the covariate x through the design used. Then by knowing the meanings of the parameters in the model, we can postulate hypotheses concerning the parameters and test these hypotheses by fitting complete and reduced models.

16.7 Exercises

16.2 A Completely Randomized Design with One Covariate

Basic

16.1 A researcher designs a study to evaluate three dietary supplements that are reputed to lower the systolic blood pressure reading for people who have high blood pressure. A inert supplement is included to evaluate the placebo effect. Twenty subjects all having systolic readings higher than 160 mmHg are randomly assigned to each of the supplements and to the control. The researcher is concerned with the disparity in age of the 80 subjects (20–60 years old) and thus wants to include the effect of age in the model also. Write a general linear model in which the response variable y, the change in systolic blood pressure after 6 months of treatment, is linearly related to the age of the subject A for each of the three supplements and the placebo. From previous studies, the researcher determines that the relationship between the reduction in blood pressure readings and age may be substantially different for the three supplements and the placebo. Identify all the parameters in your model.

Basic

16.2 Refer to Exercise 16.1. For each of the following situations, display the expected change in blood pressure for each of the four treatments (three supplements and placebo) in terms of your model parameters.

a. The four treatment lines are not parallel.
b. The four treatment lines are parallel but do not coincide.
c. The four treatment lines coincide.

Basic **16.3** Refer to Exercise 16.1. Suppose you failed to reject the hypothesis that the four treatment lines are parallel, that is, the data indicates that the four slopes are equal.

a. Describe how would you test for differences in the adjusted treatment means. Make sure to include all necessary models.
b. Provide the form of the estimated mean response for supplement 1 for subjects of age 45 years.

Basic **16.4** Refer to Exercise 16.1. After collecting the data and testing for parallelism of the four lines relating the reduction in blood pressure to the age of the subject, the researcher finds that there is significant evidence that the lines are not parallel.

a. Describe how would you test for differences in the adjusted treatment means?
b. Provide the form of the estimated mean response for supplement 1 for subjects of age 45 years.
c. The researcher wants to evaluate the supplements for subjects of age 80 years or older. What problems may she encounter using the data in her current study, if any?

Med. **16.5** A study was designed to evaluate treatments for hypertension. The researchers were concerned that whether the patient smoked might impact the effectiveness of the treatments, so they also recorded the number of cigarettes smoked daily by the patients. After 1 month on the treatment, the treating doctors assigned each patient an index based on blood pressure, cholesterol level, and amount of exercise, which reflected the patient's risk of cardiovascular disease (CVD). The index ranged from 0 to 100, with the higher values indicating a greater risk of CVD. The data are presented here with the following notation: RISK = risk index for CVD, NOCIG = number of cigarettes smoked daily, C = standard treatment, I = new treatment 1, II = new treatment 2.

Patient	RISK	NOCIG	Treatment	Patient	RISK	NOCIG	Treatment
1	22	0	C	16	42	9	I
2	26	2	C	17	50	12	I
3	49	6	C	18	54	13	I
4	67	8	C	19	70	17	I
5	72	12	C	20	82	25	I
6	19	0	C	21	12	0	II
7	28	2	C	22	14	0	II
8	97	20	C	23	17	2	II
9	88	18	C	24	29	5	II
10	30	3	C	25	37	7	II
11	7	0	I	26	45	9	II
12	9	0	I	27	53	11	II
13	14	3	I	28	81	18	II
14	18	4	I	29	93	21	II
15	30	7	I	30	94	23	II

a. Write a model for the above experiment. Make sure to identify all variables and parameters in your model.
b. Provide a scatterplot of the data with regression lines that would allow a visual assessment of whether there is a significant relationship between the CVD risk index and the number of cigarettes smoked.
c. From your scatterplot in part (b), do the three lines appear to have similar slopes?

16.6 Refer to Exercise 16.5.

- **a.** Test the hypothesis that the relationships between risk index and number of cigarettes have equal slopes for the three treatments at the $\alpha = .05$ level.
- **b.** Does there appear to be a difference in the mean risk index for the three treatments?
- **c.** Are the necessary conditions for conducting the tests of hypotheses in parts (a) and (b) satisfied with this data set?

16.3 The Extrapolation Problem

Bus. **16.7** The marketing division of a major food store chain designed the following study to evaluate three different promotions for its low-fat breakfast cereals. The promotions are as follows:

Promotion A: three boxes bundled and sold for the price of two boxes
Promotion B: a mailed-in rebate of $1 for the purchase of a mega-sized box
Promotion C: a reduction of $.50 on the price for a mega-sized box

The company wants to determine which of the three promotions produces the largest average increase in sales. Thirty stores were selected for participation in the 1-month promotion period, with 10 stores randomly assigned to one of the three promotions. The company collected data on the increase in sales (y, in hundreds of units sold) and the average monthly sales for the 12 months prior to the promotion (x, in hundreds of units). The data are given here.

	Promotion A		Promotion B		Promotion C	
Store	y	x	y	x	y	x
1	35.7	18	5.6	25	17.5	34
2	36.0	22	6.1	27	17.9	36
3	36.3	24	7.2	29	17.1	38
4	35.8	25	8.2	32	18.6	41
5	35.1	19	8.2	31	21.0	42
6	37.0	22	7.9	28	17.7	39
7	37.5	24	9.5	34	22.7	46
8	34.0	18	11.1	33	17.1	37
9	37.8	24	10.0	31	19.8	39
10	37.9	23	10.9	35	19.0	43

- **a.** Write a model for this experiment. Make sure to identify all variables and parameters in your model.
- **b.** Provide a scatterplot of the data with regression lines that would allow a visual assessment of whether there is a significant relationship between the increase in sales and the average monthly sales figures.
- **c.** From your scatterplot in part (b), do the lines associated with the three promotions appear to have similar slopes?

16.8 Refer to Exercise 16.7.

- **a.** Test the hypothesis that the relationships between increase in sales and average monthly sales have equal slopes for the three promotions at the $\alpha = .05$ level.
- **b.** Does there appear to be a difference in the increase in sales for the three promotions?
- **c.** Are the necessary conditions for conducting the tests of hypotheses in parts (a) and (b) satisfied with this data set?

16.9 Refer to Exercise 16.7.

- **a.** After carefully examining the plots and data, do you see any problems associated with the inferences made in Exercise 16.8? Justify your answer.
- **b.** If your answer in part (a) is yes, how would you redesign the study to overcome these problems?

16.4 Multiple Covariates and More Complicated Designs

Basic **16.10** In a study of allergic reactions to genetically engineered foods (GEFs), a nutritionist designed a study in which 20 subjects were exposed to five different GEFs. The order in which the subjects were exposed to the five GEFs was randomized, and there was an appropriate washout time between exposures. Let y be a measure of the allergic reaction to the exposures. The nutritionist was concerned that the subjects had very different diets in their normal habits. Thus, she devised an index, D, that measured the diversity in a subject's diet, with large values of D indicating a widely diverse diet. After running the experhnent, the nutritionist plotted the data, and the scatterplot indicated a straight-line relationship between y and D.

- **a.** Write a model for this experiment that allows a different slope for each of the five GEFs. Make sure to identify all variables and parameters in your model.
- **b.** Indicate how you would test for parallelism among the five lines. What are the degrees of freedom of the F test for parallelisn?
- **c.** Indicate how you would perform a test for differences in the mean allergic reactions to the five GEFs after adjusting for the relationship between the allergic reactions and the difference in diet diversity as measured by D.

Basic **16.11** Refer to Exercise 16.10.

- **a.** Write a model that allows a second-order relationship between y and D.
- **b.** How would you test for parallelism of the second-order model? Include the research hypothesis in terms of the model parameters and the form of the F test.
- **c.** What are the degrees of freedom of the F test for parallelism?
- **d.** Indicate how you would perform a test for the effects of treatments adjusted for the covariate.

Bio. **16.12** The seafood industry is constantly experimenting with different methods for maintaining the quality of its product during storage. One such method used in the shrimp industry is ice glazing, where the shrimp are immersed in a salt-sugar solution at a low temperature, resulting in a thin layer of ice forming on the shrimp. The coating will hopefully limit the deterioration in the quality of shrimp if there is a deviation from the required storage temperature. An experiment was designed to study the effect of the length of time the shrimp were immersed in a container of cold water, the method by which the ice glaze is applied. The immersion times (IMTs) were 5, 10, 15, 20, and 25 seconds at a standard storage temperature of –25°C. To help control for the variation in shrimp characteristics, 5 shrimp were randomly selected from each of six batches of shrimp. The shrimp were then randomly assigned to one of the immersion times. A measure of the spoilage in frozen shrimp is the total volatile base nitrogen (TVBN) level in the shrimp. The TVBN level of each of the 30 shrimp was measured at the end of 135 days in storage. Previous studies have indicated that glycogen levels in the shrimp may have an effect on the development of spoilage over longer periods of storage. Thus, the glycogen levels (GL) in the 30 shrimp were measured prior to applying the ice glaze. The data from the study are given in the following table.

Batch	IMT	GL	DVBN	Batch	IMT	GL	DVBN
1	5	2.62	17.53	4	5	2.45	17.85
	10	2.61	17.37		10	2.47	17.74
	15	2.69	17.40		15	2.42	17.69
	20	2.65	17.34		20	2.44	17.67
	25	2.61	17.41		25	2.41	17.66
2	5	3.63	16.82	5	5	2.37	17.92
	10	3.64	16.70		10	2.36	17.91
	15	3.54	16.79		15	2.32	17.87
	20	3.59	16.57		20	2.39	17.60
	25	3.61	16.48		25	2.43	17.51
3	5	2.83	17.44	6	5	3.07	17.33
	10	2.76	17.55		10	3.11	17.20
	15	2.85	17.38		15	3.09	17.15
	20	2.84	17.26		20	3.06	17.24
	25	2.85	17.13		25	3.13	17.01

a. Write a linear model relating the DVBN levels in the shrimp to the immersion times, with an adjustment for the GL levels in the shrimp prior to ice glazing.
b. Using a scatterplot, does there appear to be straight-line relationship between the DVBN and GL levels in the shrimp? Make sure to take into account the six batches and IMT values.
c. Test the research hypothesis that there is a difference in the slopes relating DVBN to GL across the five values of IMT.

Bio. **16.13** Refer to Exercise 16.12.

a. Based on your results in Exercise 16.12, test for differences in the mean level of DVBN across the five levels of immersion times.
b. Estimate the mean level of DVBN in shrimp having an immersion time of 20 seconds and a glycogen level of 3.0.

Bio. **16.14** Refer to Exercise 16.12.

a. Test for differences in the mean levels of DVBN across the five levels of immersion times without adjusting for glycogen level.
b. Estimate the mean level of DVBN in shrimp having an immersion time of 20 seconds.
c. Based on your results in Exercise 16.13 and parts (a) and (b) of this exercise, did adjusting for the glycogen level have any impact on your results?

Supplementary Exercises

Med. **16.15** An investigator studied the effects of three different antidepressants (A, B, and C) on patient ratings of depression. To do this, patients were stratified into six age–gender combinations. From a random sample of three patients from each stratum, the experimenter randomly allocated the three antidepressants. On the day the study was to be initiated, a baseline (pretreatment) depression scale rating was obtained from each patient. The assigned therapy was then administered and maintained for 1 week. At that time, a second rating (posttreatment) was obtained from each patient. The pre- and posttreatment ratings appear next (a higher score indicates more depression).

Block	Gender	Age (years)	Pretreatment A	Pretreatment B	Pretreatment C	Posttreatment A	Posttreatment B	Posttreatment C
1	F	<20	48	36	31	21	25	17
2	F	20–40	43	31	28	22	21	19
3	F	>40	44	35	29	18	24	18
4	M	<20	42	38	29	26	20	17
5	M	20–40	37	34	28	21	24	15
6	M	>40	41	36	26	18	24	19

a. Identify the experimental design.
b. Write a first-order model relating the posttreatment response y to the pretreatment rating x_1 for each treatment.

16.16 Refer to Exercise 16.15.

a. Use a computer program to fit the model of part (b) of Exercise 16.15. Use $\alpha = .05$.
b. Test for parallelism of the lines.
c. Assuming that the lines are parallel, test for differences in treatment means adjusted for the covariate. Use $\alpha = .05$.

16.17 Refer to Exercises 16.15 and 16.16.

a. Assuming parallelism of the response lines, perform a test for block differences adjusted for the covariate. Use $\alpha = .05$.
b. How might you partition the block sum of squares into five meaningful single-degree-of-freedom sums of squares?
c. Write a model and perform the tests suggested in part (b). Use $\alpha = .05$.

Soc.

16.18 A study was designed to evaluate whether socioeconomic factors had an effect on verbalization skills of young children. Four socioeconomic classes were defined, and 20 children under the age of six were selected for the study. The research hypothesis was that the mean verbalization skills would be different for the four classes. The researchers determined that for young children there may be significant gains in verbalization skills over only a few months. Thus, they decided to record the exact age (in months) of each child. The verbalization skills (measured by testing) were determined for each child. The data are given here.

Socioeconomic Class 1		2		3		4	
Age (months)	Verbal Skill	Age (months)	Verbal Skill	Age (months)	Verbal Skill	Age (months)	Verbal Skill
40	26.2	20	20.8	54	34.3	27	33.1
37	37.5	65	39.0	27	25.1	36	37.1
30	19.6	51	34.3	25	27.0	23	47.3
61	43.2	56	39.4	44	29.1	31	47.3
41	32.4	16	23.7	31	33.3	48	53.7
21	23.5	29	23.8	39	38.4	48	59.6
18	15.6	20	37.2	25	14.9	16	36.0
36	18.5	20	33.0	18	38.7	32	41.2
16	23.6	17	21.9	17	32.7	31	44.2
41	21.0	35	36.1	22	34.0	24	48.9

(coutinues)

(coutinues)

Socioeconomic Class							
1		2		3		4	
Age (months)	Verbal Skill	Age (months)	Verbal Skill	Age (months)	Verbal Skill	Age (months)	Verbal Skill
19	11.9	25	31.7	24	23.8	20	53.0
30	10.2	21	37.6	28	13.3	26	42.8
26	29.8	27	26.0	23	32.4	24	50.8
28	20.6	25	20.3	17	36.2	33	42.1
16	13.5	25	32.6	26	33.7	21	42.6
28	17.2	28	25.8	23	29.2	25	45.0
19	29.3	33	21.2	26	33.2	37	59.8
34	25.6	16	36.3	35	28.5	36	37.9
20	25.6	22	34.2	31	31.4	19	38.9
18	18.4	23	17.7	37	39.2	34	45.0

a. Plot the sample data. Do verbalization skill and age appear to be linearly related for each of the four groups?

b. Write a first-order model relating verbalization skill to age with a separate line for each socioeconomic group.

16.19 Refer to Exercise 16.18.

a. Test whether the equations relating verbalization skill to age for the four socioeconomic groups are parallel lines.

b. Are there significant differences in the mean verbalization scores for the four groups? Test this hypothesis using $\alpha = .05$.

c. Place 95% confidence intervals on the mean adjusted verbalization scores for the four groups.

Engin. **16.20** A process engineer designed a study to evaluate the differences in the mean film thicknesses of a coating placed on silicon wafers using three different coating processes. From a batch of 30 homogeneous silicon wafers, 10 wafers are randomly assigned to each of the three processes. The film thickness (y) and the temperature (x) in the lab during the coating process are recorded on each wafer. The engineer is concerned that fluctuations in the lab temperature have an effect on the thickness of the coating. The data are given here.

Wafer	x	y	Process	Wafer	x	y	Process
1	26	100	P1	16	35	159	P2
2	35	150	P1	17	26	126	P2
3	28	106	P1	18	30	141	P2
4	31	95	P1	19	32	147	P2
5	29	113	P1	20	31	143	P2
6	34	144	P1	21	37	124	P3
7	30	114	P1	22	31	95	P3
8	27	97	P1	23	34	120	P3
9	32	128	P1	24	27	86	P3
10	33	132	P1	25	28	98	P3
11	24	118	P2	26	25	81	P3
12	28	134	P2	27	29	96	P3
13	29	138	P2	28	30	99	P3
14	32	147	P2	29	35	118	P3
15	36	165	P2	30	32	107	P3

a. Plot the thickness of the coating versus the temperature in the lab.
b. Do the thickness and temperature appear to be linearly related for each of the three processes?
c. Write a model relating the thickness of the coating to the coating process with adjustments for the temperature in the lab during coating.
d. Use a computer program to fit the model in part (c).

16.21 Refer to Exercise 16.20.
a. Test whether the three equations relating thickness to temperature are parallel.
b. Test at the $\alpha = .05$ level if there is a significant difference in the mean thicknesses of the coating from the three processes after adjusting for the temperature in the lab.
c. Place 95% confidence intervals on the mean adjusted thicknesses of the coating for the three processes.

16.22 Refer to Exercise 16.21.
a. Test at the $\alpha = .05$ level if there is a significant difference in the mean thicknesses of the coating from the three processes without taking into account the temperature in the lab.
b. Are your conclusions from part (a) consistent with your conclusions from Exercise 16.21? Explain your answer.

Bio. **16.23** ***Pyke et al. (2001)*** describe a study that deals with the floristic composition of lowland tropical forest in the watershed of the Panama Canal. The following variables were measured on 45 plots in five regions: Stems—number of tree stems; Species—number of tree species; Fisher's alpha and Shannon index (H), which are measures of biodiversity of the foliage; Topography—1 = level terrain, 2 = sloping, 3 = irregular; Age—1 = secondary forest, 2 = mature secondary, 3 = old growth, primary forest; Ppt = annual precipitation (mm); PptDry = dry season precipitation (mm).

Region	Plot	Stems	Species	FisherAlpha	ShannonH	Topography	Age	Ppt	PptDry
1	1	400	84	31.41	3.13	2	3	2,589	697
1	2	409	90	35.67	3.90	2	3	2,586	696
1	3	365	98	40.91	3.82	2	3	2,579	695
1	4	450	87	33.92	4.06	2	3	2,572	693
1	5	364	93	32.80	3.43	2	3	2,594	697
1	6	480	75	22.67	3.62	2	3	2,589	697
1	7	457	78	28.81	3.89	1	2	2,529	667
1	8	467	75	25.73	3.70	3	3	2,516	647
1	9	461	74	23.59	3.02	3	2	2,497	618
1	10	429	60	16.15	3.89	3	2	2,576	659
1	11	519	92	38.53	3.66	1	3	2,535	652
2	12	380	50	17.48	3.95	3	1	1,888	524
2	13	560	49	17.96	3.54	3	1	1,890	525
2	14	503	57	23.57	3.91	3	1	1,892	525
2	15	403	58	21.49	3.65	3	1	1,887	524
2	16	172	65	23.33	3.79	3	1	1,969	568
2	17	186	64	24.83	3.98	3	1	2,096	638
3	18	449	63	21.02	3.43	1	2	2,993	720
3	19	520	84	32.03	3.33	3	3	3,072	780
3	20	647	74	28.02	2.44	3	1	3,007	811
3	21	381	94	54.76	3.93	3	1	3,000	810
3	22	409	88	31.16	3.76	3	2	3,026	792
3	23	408	81	26.63	3.97	3	2	3,026	792
3	24	407	65	20.2	3.74	3	2	3,028	792

(coutinues)

(coutinued)

Region	Plot	Stems	Species	Fisher Alpha	ShannonH	Topography	Age	Ppt	PptDry
3	25	526	75	23.92	3.42	3	2	3,030	793
3	26	597	70	17.40	2.65	3	1	3,032	793
4	27	531	71	26.33	3.75	2	2	2,414	621
4	28	484	78	26.41	3.55	3	2	2,394	612
4	29	526	93	39.27	3.81	3	1	2,438	638
4	30	954	94	32.32	3.06	3	3	2,456	635
4	31	424	107	41.60	3.53	3	3	2,889	924
4	32	534	91	34.13	3.70	1	3	2,455	646
4	33	405	90	33.17	3.76	3	3	2,502	707
4	34	508	63	19.73	3.37	3	3	2,471	679
4	35	579	86	32.37	2.70	1	2	2,511	645
4	36	557	89	30.92	2.95	3	1	2,688	743
4	37	593	90	31.01	3.80	3	1	2,658	737
5	38	485	78	28.73	3.41	3	1	2,411	662
5	39	393	75	24.30	3.45	3	1	2,514	722
5	40	408	60	16.82	3.33	3	2	2,248	585
5	41	355	60	17.07	3.33	3	2	2,280	602
5	42	302	84	26.26	3.16	3	2	2,334	641
5	43	466	76	25.30	4.55	3	2	2,252	591
5	44	148	61	20.21	4.40	3	1	2,305	681
5	45	191	62	20.35	4.11	3	1	2,294	668

a. Is there significant evidence of a difference among the three age classifications of the forests relative to their biodiversity as measured by Fisher's alpha. Use annual precipitation to adjust for differences in the five regions.

b. Provide a grouping of the three age classifications based on their adjusted mean Fisher's alpha.

c. Using residual plots, evaluate whether the conditions needed to properly answer parts (a) and (b) are valid for this data set.

Bio. **16.24** Refer to Exercise 16.23.

a. Is there significant evidence of a difference among the three topography classifications of the forests relative to their biodiversity as measured by Fisher's alpha. Use annual precipitation to adjust for differences in the five regions.

b. Provide a grouping of the three topography classifications based on their adjusted mean Fisher's alpha.

c. Using residual plots, evaluate whether the conditions needed to properly answer parts (a) and (b) are valid for this data set.

Bio. **16.25** Refer to Exercise 16.23. Researchers have opined that in many forests the Shannon index is a more complete measure of biodiversity than Fisher's alpha. Repeat Exercises 16.23 and 16.24 using Shannon's index in place of Fisher's alpha as a measure of biodiversity. Are there any differences in your conclusions?

Bio. **16.26** Refer to Exercise 16.23. Biologists have noted that in many environments the annual precipitation is not the crucial factor in the survival of many types of foliage; rather, it is the amount of precipitation during the dry season. Repeat Exercises 16.23 and 16.24 using the dry season precipitation in place of the annual precipitation. Are there any differences in your conclusions?

Bio. **16.27** In Exercises 16.23–16.26, conclusions were drawn separately for the effects of age and topography on biodiversity of the forests. Using the separate analyses, it is not possible to determine if the effects of age are consistent for the three types of topography.

a. Write a model relating Fisher's alpha to the main effects and interaction of age and topography, using annual precipitation as the adjustment for differences in the five regions.

b. If possible, conduct an analysis of the combined effects of age and topography on the biodiversity of the forests with an adjustment for annual precipitation differences across the five regions.

c. If it was not possible to conduct the analysis requested in part (b), modify the factors age and topography in such a manner that an analysis can be conducted.

CHAPTER 17

Analysis of Variance for Some Fixed-, Random-, and Mixed-Effects Models

17.1 Introduction and Abstract of Research Study

The experiments and studies we encountered in previous chapters all involved experimental factors and treatments in which the researchers selected particular levels of the treatments for study. These were the only levels for which inferences would be made from the experimental data. The case study in Chapter 16 involved three new cultivars, and these were the only cultivars of interest to the researchers. In this experiment, the only populations of interest were the three populations of greens speeds for the three cultivars.

If the USGA decided it was necessary to repeat the experiments in order to verify the mean speeds obtained in the original experiment, the three cultivars could be planted on another set of greens and the experiments duplicated. In a study or experiment involving factors having a predetermined set of levels, the model used to examine the variability in the response variable is referred to as a **fixed-effects** model. The inferences from these models are restricted to the particular set of treatment levels used in the study.

fixed-effects

DEFINITION 17.1

In a **fixed-effects model** for an experiment, all the factors in the experiment have a predetermined set of levels, and the only inferences are for the levels of the factors actually used in the experiment.

variance components

The major interest in some studies is to identify factors that are sources of variability in the response variable. In product improvement studies, the quality control engineer attempts to determine which factors in the production process are the major sources of variability, referred to as **variance components,** and to estimate the contribution of each of these sources of variability to the overall variability in the product. When the levels of the factors to be used in the experiment are randomly selected from a population of possible levels, the model used to relate the response variable to the levels of the factors is referred to as a **random-effects** model. The inferences from these models are generalized to the population of levels from the levels used in the experiment, which were randomly selected. In a product improvement study, one of the common sources of variability is the operator of the process. The company may have hundreds of operators, but only five or six will be randomly selected to participate in the study. However, the quality engineer is interested in the performance of all operators, not only the operators that are involved in the study.

random-effects

DEFINITION 17.2

In a **random-effects model** for an experiment, the levels of factors used in the experiment are randomly selected from a population of possible levels. The inferences from the data in the experiment are for all levels of the factors in the population from which the levels were selected and not only the levels used in the experiment.

Many studies will involve factors having a predetermined set of levels and factors in which the levels used in the study are randomly selected from a population of levels. The blocks in a randomized complete block design might represent a random sample of b plots of land taken from a population of plots in an agricultural research facility. Then the effects due to the blocks are considered to be random effects. Suppose the treatments are four new varieties of soybeans that have been developed to be resistant to a specific virus. The levels of the treatment are fixed because these are the only varieties of interest to the researchers, whereas the levels of the plots of land are random because the researchers are interested in the effects of these treatments not only on these plots of land but also on a wide range of plots of land. When some of the factors to be used in the experiment have levels randomly selected from a population of possible levels and other factors have predetermined levels, the model used to relate the response variable to the levels of the factors is referred to as a **mixed-effects** model.

DEFINITION 17.3

In a **mixed-effects model** for an experiment, the levels of some of the factors used in the experiment are randomly selected from a population of possible levels, whereas the levels of the other factors in the experiment are predetermined. The inferences from the data in the experiment concerning factors with fixed levels are only for the levels of the factors used in the experiment, whereas inferences concerning factors with randomly selected levels are for all levels of the factors in the population from which the levels were selected.

In this chapter, we will consider various random-effects and mixed-effects models. For each model, we will indicate the appropriate analysis of variance and show how to estimate all relevant components of variance. The following research study will describe a mixed-effects experiment.

Abstract of Research Study: Factors Affecting Pressure Drops Across Expansion Joints

A major problem in power plants is that of pressure drops across expansion joints in electric turbines. The process engineer wants to design a study to identify the factors that are most likely to influence the pressure drop readings. Once these factors are identified and the most crucial factors are determined by the sizes of their contributions to the pressure drops across the expansion joints during the study, the engineer can make design changes in the process or alter the method by which the operators of the process are trained. These types of changes may be expensive or time consuming, so the engineer wants to be certain which factors will have the greatest impact on reducing the pressure drops.

The factors selected for study are the gas temperature on the inlet side of the joint and the type of pressure gauge used by the operator. The engineer decides that a design with a factorial treatment structure is required to determine which of these factors has the greatest effect on the pressure drop. Three temperatures that cover the feasible range for operation of the turbine are 15°C, 25°C, and 35°C. There are hundreds of different types of pressure gauges used to monitor the pressure in the lines. Four types of gauges are randomly selected from the list of possible gauges for use in the study. In order to obtain a precise estimate of the mean pressure drop for each of the 12 combinations of temperature and type of gauge, it was decided to obtain six replications of each of 12 treatments. The data from the 72 experimental runs are given in Table 17.1.

In order to determine if the observed differences displayed in Table 17.1 are more than just random variation, we will develop models and analysis techniques in the remainder of this chapter to enable us to identify which factors make the greatest contribution to the overall variation in the pressure drop across the expansion joints.

TABLE 17.1 Pressure drop across expansion joints

	Temperature											
	15°C				25°C				35°C			
	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4
	40	43	42	47	57	49	44	36	35	41	42	41
	40	34	35	47	57	43	45	49	35	43	41	44
	37	38	35	40	65	51	49	38	35	44	34	35
	47	42	41	36	67	49	45	45	46	36	35	46
	42	39	43	41	63	45	46	38	41	42	39	44
	41	35	36	47	59	43	43	42	42	41	36	46
Mean	41.17	38.50	38.67	43.00	61.33	46.67	45.33	41.33	39.00	41.17	37.83	42.67

17.2 A One-Factor Experiment with Random Treatment Effects

The best way to illustrate the difference between the fixed- and random-effects models for a one-factor experiment is by an example. Suppose we want to compare readings made on the intensities of the electrostatic discharges of lightning at three different tracking stations within a 20-mile radius of the central computing facilities of a university. If these three tracking stations are the only feasible tracking stations for such an operation and inferences are to be about these stations only, then we could write the **fixed-effects model** as

fixed-effects model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \text{with } \mu_i = E(y_{ij}) = \mu + \tau_i$$

where y_{ij} is the jth observation at tracking station i ($i = 1, 2, 3$), μ is an overall mean, and τ_i is a fixed effect due to tracking station i. For both of these models, ε is assumed to be normally distributed with mean 0 and variance σ^2.

Suppose, however, that rather than being concerned about only these three tracking stations, we consider these stations as a random sample of three taken from the many possible locations for tracking stations. Inferences would now relate not only to what happened at the sampled locations but also to what might happen at other possible locations for tracking stations. A model that can account for this difference in interpretation is the **random-effects model:**

random-effects model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \text{with } \mu_i = E(y_{ij}) = \mu$$

Although the model looks the same as the previous fixed-effects model, some of the **assumptions** are different.

assumptions

1. μ is still an overall mean, which is an unknown constant.
2. τ_i is a random effect due to the ith tracking station. We assume that τ_i is normally distributed with mean 0 and variance σ_τ^2.
3. The τ_is are independent.
4. As before, ε_{ij} is normally distributed with mean 0 and variance σ_ε^2.
5. The ε_{ij}s are independent.
6. The random components τ_i and ε_{ij} are independent.

The difference between the fixed-effects model and the random-effects model can be illustrated by supposing we were to repeat the experiment. For the fixed-effects model, we would use the same three tracking stations, so it would make sense to make inferences about the mean intensities or differences in mean intensities at these three locations. However, for the random-effects model, we would take another random sample of three tracking stations (i.e., take another sample of three τ_is). Now, rather than concentrating on the effect of a particular group of three τ_is from one experiment, we would examine the variability of the population of all possible τ_i values. This will be illustrated using the analysis of variance table given in Table 17.2.

TABLE 17.2
An AOV table for a one-factor experiment: fixed or random model

				EMS	
Source	**SS**	**df**	**MS**	**Fixed Effects**	**Random Effects**
Treatments	SST	$t - 1$	MST	$\sigma_\varepsilon^2 + n\theta_\tau$	$\sigma_\varepsilon^2 + n\sigma_\tau^2$
Error	SSE	$t(n - 1)$	MSE	σ_ε^2	σ_ε^2
Totals	TSS	$tn - 1$			

EMS

The analysis of variance table is the same for a fixed- or random-effects model except that the **expected mean squares (EMS)** columns are different. You will recall that this column was not used in our tables in Chapters 14 and 15 because all mean squares except MSE had an expectation under the alternative hypothesis equal to σ_ε^2 plus a positive constant, which depended on the parameters under test. In general, with t treatments (tracking stations) and n observations per treatment,

AOV table

the **AOV table** would appear as shown in Table 17.2. For the fixed-effects model, θ_τ is a positive function of the constants τ_i, whereas σ_τ^2 represents the variance of the population of τ_i values for the random-effects model. Referring to our example, a **test for the equality of the mean** intensities at the three tracking stations in the fixed-effects model is (from Chapter 14)

test for means

H_0: $\mu_1 = \mu_2 = \mu_3$
H_a: At least one μ_i is different from the rest.

In terms of model parameters:

H_0: $\tau_1 = \tau_2 = \tau_3 = 0$
H_a: At least one τ_i is different from 0.
T.S.: $F = \text{MST}/\text{MSE}$, based on $\text{df}_1 = t - 1$ and $\text{df}_2 = t(n-1)$

test for σ_τ^2

A **test concerning the variability for the population of τ** values in the random-effects model makes use of the same test statistic. The null hypothesis and alternative hypothesis are

H_0: $\sigma_\tau^2 = 0$
H_a: $\sigma_\tau^2 > 0$
T.S.: $F = \text{MST}/\text{MSE}$, based on $\text{df}_1 = t - 1$ and $\text{df}_2 = t(n-1)$

Because we assumed that the τ_is sampled were selected from a normal population with mean 0 and variance σ_τ^2, the null hypothesis states that the τ_is were drawn from a normal population with mean 0 and variance 0; that is, all τ values in the population are equal to 0.

Thus, although the forms of the null hypotheses are different for the two models, the meanings attached to them are very similar. For the fixed-effects model, we are assuming that the three τs in the model (which are the only τs) are identically 0, whereas in the random-effects model, the null hypothesis leads us to assume that the sampled τs, as well as all other τs in the population, are 0.

The alternative hypotheses are also similar. In the fixed-effects model, we are assuming that at least one of the τs is different from the rest; that is, there is some variability among the set of τs. For the random-effects model, the alternative hypothesis is that $\sigma_\tau^2 > 0$; that is, not all τ values in the population are the same.

In a random-effects model with a single factor, the response variable has a mean value and variance given by

$$E(y_{ij}) = \mu \quad \text{and} \quad \sigma_y^2 = \text{Var}(y_{ij}) = \sigma_\tau^2 + \sigma_\varepsilon^2$$

Thus, in many random-effects experiments, we want to determine the size of σ_τ^2 relative to that of σ_ε^2 in order to assess the size of the treatment effect relative to the overall variability in the response variable. Because we do not know σ_τ^2 or σ_ε^2,

AOV moment matching

we can form estimates of these terms by using the idea of **AOV moment matching** estimators. From Table 17.3, we see that MST has an expected mean square of $\sigma_\varepsilon^2 + n\sigma_\tau^2$ and MSE has an expected mean square of σ_ε^2.

TABLE 17.3
AOV table with expected mean squares

Source	MS	EMS
Treatments	MST	$\sigma_\varepsilon^2 + n\sigma_\tau^2$
Error	MSE	σ_ε^2

When we equate the sample mean square to its expected value and solve for the population variance, we get

$$\hat{\sigma}_\varepsilon^2 = \text{MSE} \text{ and } \hat{\sigma}_\tau^2 = (\text{MST} - \text{MSE})/n$$

As a result, we have $\hat{\sigma}_y^2 = \hat{\sigma}_\tau^2 + \hat{\sigma}_\varepsilon^2$. The variance in the response variable can thus be proportionally allocated to the two sources of variability, the treatment and the experimental error, shown in Table 17.4.

TABLE 17.4
Proportional allocation of total variability in the response variable

Source of Variance	Estimator	Proportion of Total
Treatment	$\hat{\sigma}_\tau^2 = (\text{MST} - \text{MSE})/n$	$\hat{\sigma}_\tau^2/\hat{\sigma}_y^2$
Error	$\hat{\sigma}_\varepsilon^2 = \text{MSE}$	$\hat{\sigma}_\varepsilon^2/\hat{\sigma}_y^2$
Total	$\hat{\sigma}_y^2 = \hat{\sigma}_\tau^2 + \hat{\sigma}_\varepsilon^2$	1.0

It might also be of interest to the researchers to estimate the mean value for the response variable, μ. The point estimator of μ and its estimated standard error are given by

$$\hat{\mu} = \bar{y}_{..} \text{ and } \text{SE}(\hat{\mu}) = \sqrt{\text{MST}/tn}$$

We can then construct a $100(1 - \alpha)\%$ confidence interval for μ, as given here.

$$\hat{\mu} \pm t_{\alpha/2,\, \text{df}_{\text{TRT}}} \text{SE}(\hat{\mu}) = \bar{y}_{..} \pm t_{\alpha/2,\, t-1}\sqrt{\text{MST}/tn}$$

EXAMPLE 17.1

Consider the problem we used to illustrate a one-factor experiment with random treatment effects. Two graduate students working for a professor in electrical engineering have been funded to record lightning discharge intensities (intensities of the electrical field) at three tracking stations. Because of the high frequency of thunderstorms in the summer months (in Florida, storms occur on 80 or more days per year), the graduate students were to choose a point at random on a map of the 20-mile-radius region and assemble their tracking equipment (provided they could get permission of the property owner). Each day from 8 A.M. to 5 P.M., they were to monitor their instruments until the maximum intensity had been recorded for five separate storms. They then repeated the process separately at the two other locations chosen at random. The sample data (in volts per meter) appear in Table 17.5.

TABLE 17.5
Lightning discharge intensities (in volts per meter)

Tracking Station	Intensities					Mean
1	20	1,050	3,200	5,600	50	1,984
2	4,300	70	2,560	3,650	80	2,132
3	100	7,700	8,500	2,960	3,340	4,520
Overall mean						2,878.67

a. Write an appropriate statistical model, defining all terms.
b. Perform an analysis of variance and interpret your results. Use $\alpha = .05$.
c. Estimate the variance components and their proportional allocations of the total variability.
d. Estimate the mean maximum daily lightning discharge intensity, and place a 95% confidence on this mean.

Solution **a.** Because the tracking stations were selected at random, we can use a single-factor random-effects model to relate maximum lightning discharge intensity, y_{ij}, to the ith station and jth day.

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad (i = 1, 2, 3; \quad j = 1, 2, \ldots, 5)$$

where μ is the mean maximum daily lightning discharge intensity, τ_i is the random effect of the ith randomly selected station, and ε_{ij} is the random effect due to all other sources of variability.

b. The formulas for computing the sums of squares for the random-effects analysis of variance are identical to the formulas used in the fixed-effects analysis of variance. Thus, we have

$$\text{SST} = n\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 = 5\{(1{,}984 - 2{,}878.67)^2 + (2{,}132 - 2{,}878.67)^2 + (4{,}520 - 2{,}878.67)^2\} = 20{,}259{,}573.3$$

$$\text{TSS} = \sum_{ij} (y_{ij} - \bar{y}_{..})^2 = (20 - 2{,}878.67)^2 + (1{,}050 - 2{,}878.67)^2 + \cdots + (3{,}340 - 2{,}878.67)^2 = 108{,}249{,}173.3$$

By subtraction,

$$\text{SSE} = \text{TSS} - \text{SST} = 108{,}249{,}173.3 - 20{,}259{,}573.3 = 87{,}989{,}600$$

We can use these calculations to construct an AOV table, as shown in Table 17.6.

TABLE 17.6
AOV table for the data of Example 17.1

Source	SS	df	MS	EMS	F
Tracking stations	20,259,573.3	2	10,129,786.65	$\sigma_\varepsilon^2 + 5\sigma_\tau^2$	1.38
Error	87,989,600.0	12	7,332,466.67	σ_ε^2	
Totals	108,249,173.3	14			

The F test for H_0: $\sigma_\tau^2 = 0$ is based on $\text{df}_1 = 2$ and $\text{df}_2 = 12$. Because the computed value of F, 1.38, does not exceed 3.89, the value in Appendix Table 8 for $\alpha = .05, \text{df}_1 = 2,$ and $\text{df}_2 = 12,$ we have insufficient evidence to indicate that there is a significant random component due to variability in intensities from tracking station to tracking station. Rather, as an electrical engineer postulated, it is probably best to work with a single tracking station because most of the variability in intensities is related to the distance of the tracking station from the point of discharge and we have no control over this source.

c. In fact, we can compute estimates of the variance components and obtain

$$\hat{\sigma}_\varepsilon^2 = 7{,}332{,}466.67 \quad \hat{\sigma}_\tau^2 = (10{,}129{,}786.65 - 7{,}332{,}466.67)/5 = 559{,}464$$

which yields

$$\hat{\sigma}_y^2 = 7{,}332{,}466.67 + 559{,}464 = 7{,}891{,}930.67$$

The proportion of the total variability due to station differences is 559,464/7,891,930.67 = .0709. Thus, only 7.1% of the variability in maximum daily lightning intensities is due to station differences.

d. We can place a 95% confidence interval on the mean maximum daily lightning intensity as given here.

$$\bar{y}_{\cdot\cdot} \pm t_{.025,2}\,\mathrm{SE}(\hat{\mu})$$

$$2{,}878.67 \pm (4.303)\sqrt{10{,}129{,}786.65/15} = 2{,}878.67 \pm 3{,}536.11$$

Thus, we are 95% confident that the mean daily maximum lightning intensity is within (0, 6,414.78). ■

17.3 Extensions of Random-Effects Models

The ideas presented for a random-effects model in a one-factor experiment can be extended to any of the block designs and factorial experiments covered in Chapters 14 and 15. Although we will not have time to cover all such situations, we will consider first a randomized block design in which the block effects and the treatment effects are random.

EXAMPLE 17.2

An experiment was designed to examine if there was a large variation in the DNA content of plaque due to the difference in the skills and training of the analysts conducting the chemical analysis. A random sample of five analysts was taken from the population of analysts certified to conduct the DNA analysis. Ten female subjects (ages 18–20) were chosen for the study. Each subject was allowed to maintain her usual diet, supplemented with 30 mg of sucrose per day. No brushing of teeth or use of mouthwash was allowed during the study. At the end of the week, plaque was scraped from the entire dentition of each subject and divided into five samples. Each of the five randomly selected analysts was then given an unmarked sample of plaque from each of the 10 subjects. An analysis for the DNA content (in micrograms) was then performed. The data are shown in Table 17.7. Identify the design and provide a model for this experiment.

TABLE 17.7
DNA concentrations in plaque (micrograms)

	Subjects										
Analyst	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean**
1	5.2	6.0	7.2	7.8	9.2	10.9	12.0	12.9	14.0	14.9	10.03
2	4.8	6.1	6.9	7.9	9.1	11.0	12.2	12.8	13.9	15.1	9.99
3	5.4	6.2	7.2	8.3	9.4	11.4	12.4	13.6	14.2	15.2	10.32
4	5.2	6.2	7.4	8.3	9.6	10.9	12.2	13.2	14.3	15.6	10.30
5	5.7	7.0	7.9	8.8	9.7	11.7	12.8	13.9	15.0	15.7	10.81
Mean	5.26	6.30	7.31	8.21	9.39	11.19	12.31	13.32	14.32	15.30	10.29

randomized block design

Solution This experimental design is recognized as a **randomized block design,** with subjects representing blocks and analysts being the treatments. The experimental units are samples of plaque scraped from the dentition of the subjects. If we assume that the 10 subjects represent a random sample from a large population of

random-effects model

possible subjects and, similarly, that the five analysts represent a random sample from a large population of possible analysts, we can write the following **random-effects model** relating DNA concentration to the two factors, analysts and subjects:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

Model Conditions:

assumptions

1. μ is an overall unknown concentration mean.
2. τ_i is a random effect due to the ith analyst. τ_i is normally distributed with mean 0 and variance σ_τ^2.
3. The τ_is are independent.
4. β_j is a random effect due to the jth subject. β_j is a normally distributed random variable with mean 0 and variance σ_β^2.
5. The β_js are independent.
6. The τ_is, β_js, and ε_{ij}s are mutually independent.

Again note the difference between assuming that the treatments and blocks are random, and assuming that they are fixed. If, for example, the five analysts chosen for the study were the only analysts of interest, we would be concerned with differences in mean DNA concentrations for these specific analysts. Now, however, treating the effect due to an analyst as a random variable, our inference will be about the population of analysts' effects. Because the mean of this normal population is assumed to be 0, we want to determine whether the variance σ_τ^2 is greater than 0. ■

The AOV table for a randomized block design with t treatments is given in Table 17.8. There are two columns for the expected mean squares. The first column is for the situation in which the treatment and block effects are fixed, and the second column is for the situation in which the treatment and block effects are random. The formulas for sum of squares block (SSB) and sum of squares treatment (SST) are identical to the formulas used when both the block and treatment effects are fixed, as were developed in Chapter 15. Likewise, the F tests are identical to the F tests for experiments having both block and treatment effects fixed. However, there is a major difference between the two models with respect to the types of inferences made from the results of the F tests. In the fixed block effects case, inferences are restricted to the levels of the blocks used just in the experiment. In the random block effects case, we are making inferences about the population of blocks from which the blocks used in the experiment were randomly selected. This provides for more general and realistic results in that the block effects often involve not only the physical entities (subjects in Example 17.2) but also differences in the environmental conditions encountered during the experiment. The differences in the inferences between fixed and random effects are reflected in the expected mean squares.

TABLE 17.8
AOV table for a randomized block design with r blocks and t treatments

				EMS	
Source	**SS**	**df**	**MS**	**Fixed TRT, BL Effects**	**Random TRT, BL Effects**
Block	SSB	$r - 1$	MSB	$\sigma_\varepsilon^2 + t\theta_\beta$	$\sigma_\varepsilon^2 + t\sigma_\beta^2$
Treatment	SST	$t - 1$	MST	$\sigma_\varepsilon^2 + r\theta_\tau$	$\sigma_\varepsilon^2 + r\sigma_\tau^2$
Error	SSE	$(r - 1)(t - 1)$	MSE	σ_ε^2	σ_ε^2
Total	TSS	$rt - 1$			

TABLE 17.9
Difference in test procedures for treatments

Fixed-Effects Model	Random-Effects Model
H_0: $\tau_1 = \tau_2 = \cdots = \tau_t = 0$	H_0: $\sigma_\tau^2 = 0$
H_a: At least one of the τ_is differs from the rest.	H_a: $\sigma_\tau^2 > 0$
T.S.: $F = \dfrac{\text{MSA}}{\text{MSE}}$	T.S.: $F = \dfrac{\text{MSA}}{\text{MSE}}$
R.R.: Based on $df_1 = t - 1$ and $df_2 = (t-1)(b-1)$	R.R.: Same

The computation of sums of squares and mean squares would proceed exactly as shown in Chapter 15. The difference in test procedures is illustrated in Table 17.9 for treatments.

Rather than proceeding with an example at this point, we will discuss a random-effects model for a factorial treatment structure with $n > 1$ observations at each factor–level combination. Then we will illustrate the test procedure.

$a \times b$ factorial treatment structure, $n > 1$

In Chapter 14, we considered the fixed-effects model for an $a \times b$ factorial treatment structure in a completely randomized design with $n > 1$ observations per cell. The random-effects model for an ***a* × *b* factorial treatment structure*** would be of the same form as the corresponding model for a fixed-effects experiment, but with different assumptions:

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the response of the kth observation at the ith level of factor A and jth level of factor B; μ is the overall mean response; τ_i is the main effect of the ith level of factor A; β_j is the main effect of the jth level of factor B; $\tau\beta_{ij}$ is the interaction effect of the ith level of factor A combined with the jth level of factor B; and ε_{ijk} is the random effect. The model conditions are as follows.

Model Conditions:

1. μ is the overall mean response (an unknown population parameter).
2. τ_i is a random effect due to the ith level of factor A with τ_is independently normally distributed with mean 0 and variance σ_τ^2.
3. β_j is a random effect due to the jth level of factor B with β_js independently normally distributed with mean 0 and variance σ_β^2.
4. $\tau\beta_{ij}$ is a random effect due to the ith level of factor A combined with the jth level of factor B with $\tau\beta_{ij}$s independently normally distributed with mean 0 and variance $\sigma_{\tau\beta}^2$.
5. ε_{ijk} is the random effect due to all other factors with ε_{ijk}s independently normally distributed with mean 0 and variance σ_ε^2.
6. The τ_is, β_js, $\tau\beta_{ij}$s, and ε_{ijk}s are mutually independent.

AOV tables

The appropriate **AOV tables** for fixed- and random-effects models are shown in Table 17.10.

The appropriate tests using the AB interaction sum of squares are illustrated in Table 17.11 for the two models.

Now, unlike the one-factor experiment and the two-factor experiment without replication, the test statistics for main effects are different for the fixed- and random-effects models. In addition, for the random-effects model, the tests for σ_τ^2 and σ_β^2 can proceed even when the test on the AB interaction ($\sigma_{\tau\beta}^2$) is significant. We have seen previously that for fixed-effects models, a test for main effects

TABLE 17.10
AOV table for an $a \times b$ factorial treatment structure with n observations per cell

Source	SS	df	MS	EMS	
				Fixed Effects	Random Effects
A	SSA	$a-1$	MSA	$\sigma_\varepsilon^2 + bn\theta_\tau$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2 + bn\sigma_\tau^2$
B	SSB	$b-1$	MSB	$\sigma_\varepsilon^2 + an\theta_\beta$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2 + an\sigma_\beta^2$
AB	SSAB	$(a-1)(b-1)$	MSAB	$\sigma_\varepsilon^2 + n\theta_{\tau\beta}$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2$
Error	SSE	$ab(n-1)$	MSE	σ_ε^2	σ_ε^2
Total	TSS	$abn-1$			

TABLE 17.11
A comparison of appropriate interaction tests for fixed- and random-effects models

	Fixed-Effects Model	Random-Effects Model
H_0:	$\tau\beta_{11} = \tau\beta_{12} = \cdots = \tau\beta_{ab} = 0$	$H_0: \ \sigma_{\tau\beta}^2 = 0$
H_a:	At least one $\tau\beta_{ij}$ differs from the rest.	$H_a: \ \sigma_{\tau\beta}^2 > 0$
T.S.:	$F = \dfrac{\text{MSAB}}{\text{MSE}}$	T.S.: $F = \dfrac{\text{MSAB}}{\text{MSE}}$
R.R.:	Based on $\text{df}_1 = (a-1)(b-1)$ and $\text{df}_2 = ab(n-1)$	R.R.: Same

in the presence of a significant interaction seems to make sense only when the profile plot suggests that the interaction is "orderly." For random-effects models, we are interested in identifying the various sources of variability (e.g., $\sigma_{\tau\beta}^2$, σ_τ^2, and σ_β^2) that affect the response y. Tests for σ_τ^2 and σ_β^2 do make sense even when $\sigma_{\tau b}^2$ has been shown to be greater than zero.

For the fixed-effects model following a nonsignificant test on the AB interaction, we can test for main effects due to factors A and B by using

$$F = \frac{\text{MSA}}{\text{MSE}} \quad \text{and} \quad F = \frac{\text{MSB}}{\text{MSE}}$$

respectively. As we see from the expected mean squares column of Table 17.10, no matter what the results are for the test H_0: $\sigma_{\tau\beta}^2 = 0$, we can form an F test for the components σ_τ^2 and σ_β^2 using the test procedures shown in Table 17.12. Note that the test statistics differ from those used in the fixed-effects case, where the denominator of all F statistics is MSE.

variance components

In many experiments involving factors having random effects, we will want to estimate the **variance components** σ_τ^2, σ_β^2, $\sigma_{\tau\beta}^2$, and σ_ε^2. We can once again use the AOV moment matching estimators, which are obtained by matching the sample mean squares with the expected mean squares in the AOV table and then

TABLE 17.12
Tests for an $a \times b$ factorial treatment structure with replication: random-effects model

	Factor A	Factor B
H_0:	$\sigma_t^2 = 0$	$H_0: \ \sigma_b^2 = 0$
H_a:	$\sigma_t^2 > 0$	$H_a: \ \sigma_b^2 > 0$
T.S.:	$F = \dfrac{\text{MSA}}{\text{MSAB}}$	T.S.: $F = \dfrac{\text{MSB}}{\text{MSAB}}$
R.R.:	Based on $\text{df}_1 = (a-1)$ and $\text{df}_2 = (a-1)(b-1)$	R.R.: Based on $\text{df}_1 = (b-1)$ and $\text{df}_2 = (a-1)(b-1)$

TABLE 17.13
Proportional allocation of total variability in the response variable

Source of Variance	Estimator	Proportion of Total
Factor A	$\hat{\sigma}_\tau^2 = (\text{MSA} - \text{MSAB})/bn$	$\hat{\sigma}_\tau^2/\hat{\sigma}_y^2$
Factor B	$\hat{\sigma}_\beta^2 = (\text{MSB} - \text{MSAB})/an$	$\hat{\sigma}_\beta^2/\hat{\sigma}_y^2$
Interaction AB	$\hat{\sigma}_{\tau\beta}^2 = (\text{MSAB} - \text{MSE})/n$	$\hat{\sigma}_{\tau\beta}^2/\hat{\sigma}_y^2$
Error	$\hat{\sigma}_\varepsilon^2 = \text{MSE}$	$\hat{\sigma}_\varepsilon^2/\hat{\sigma}_y^2$
Total	$\hat{\sigma}_y^2 = \hat{\sigma}_\tau^2 + \hat{\sigma}_\beta^2 + \hat{\sigma}_{\tau\beta}^2 + \hat{\sigma}_\varepsilon^2$	1.0

solving for the individual variance components. Using the MSs and EMSs in Table 17.10, we obtain

$$\hat{\sigma}_\varepsilon^2 = \text{MSE}$$
$$\hat{\sigma}_{\tau\beta}^2 = (\text{MSAB} - \text{MSE})/n$$
$$\hat{\sigma}_\beta^2 = (\text{MSB} - \text{MSAB})/an$$

and

$$\hat{\sigma}_\tau^2 = (\text{MSA} - \text{MSAB})/bn$$

Also, from the random-effects model for two factors having randomly selected levels, we have

$$E(y_{ijk}) = \mu \text{ and } \sigma_y^2 = \sigma_\tau^2 + \sigma_\beta^2 + \sigma_{\tau\beta}^2 + \sigma_\varepsilon^2$$

Thus, we have $\hat{\sigma}_y^2 = \hat{\sigma}_\tau^2 + \hat{\sigma}_\beta^2 + \hat{\sigma}_{\tau\beta}^2 + \hat{\sigma}_\varepsilon^2$. We can then proportionally allocate the total variability $\hat{\sigma}_y^2$ into the four sources of variability: factor A, factor B, the interaction, and experimental error. See Table 17.13.

The researchers might also be interested in estimating the mean value for the response variable, μ. The point estimator of μ and its estimated standard error are given by

$$\hat{\mu} = \bar{y}_{\cdots} \quad \text{and} \quad \text{SE}(\hat{\mu}) = \sqrt{(\text{MSA} + \text{MSB} - \text{MSAB})/abn}$$

We can then construct a $100(1 - \alpha)\%$ confidence interval for μ, as given here.

$$\bar{y}_{\cdots} \pm t_{\alpha/2,\, \text{df}_{\text{Approx.}}} \sqrt{(\text{MSA} + \text{MSB} - \text{MSAB})/abn}$$

where the degrees of freedom for the t tables are obtained from the Satterthwaite approximation:

$$\text{df}_{\text{Approx.}} = \frac{(\text{MSA} + \text{MSB} - \text{MSAB})^2}{(\text{MSA})^2/(a-1) + (\text{MSB})^2/(b-1) + (\text{MSAB})^2/(a-1)(b-1)}$$

Because in most cases this value is not an integer, we take the largest integer less than or equal to $\text{df}_{\text{Approx}}$.

In some experiments, the estimates of some of the variance components may result in a negative number. Of course, by definition a variance component must be a nonnegative number; thus, we must consider alternatives whenever the sample estimator is negative.

A1. We can set the estimator equal to zero and use zero as the estimator of the variance component. However, the estimator will no longer be an unbiased estimator of the variance component.

A2. A negative estimator of a variance component may be an indication that we have elements in our model that are not appropriate for

this experiment. A more complex model may be needed for this experiment.

A3. There are alternative estimators of variance components that are mathematically beyond the level of this book. Such methods as ML or REML are available in software packages, such as SAS and R.

EXAMPLE 17.3

A consumer-product agency wants to evaluate the accuracy with which the level of calcium in a food supplement is determined. There are a large number of possible testing laboratories and a large number of chemical assays for calcium. The agency randomly selects three laboratories and three assays for use in the study. Each laboratory will use all three assays in the study. Eighteen samples containing 10 mg of calcium are prepared, and each assay–laboratory combination is randomly assigned to two samples. The determinations of calcium content are given in Table 17.14 (numbers in parentheses are averages for the assay–laboratory combinations).

TABLE 17.14
Calcium content data

	Laboratory			
Assay	**1**	**2**	**3**	**Assay Mean**
1	10.9	10.5	9.7	10.3
	10.9	9.8	10.0	
	(10.9)	(10.15)	(9.85)	
2	11.3	9.4	8.8	10.1
	11.7	10.2	9.2	
	(11.5)	(9.8)	(9.0)	
3	11.8	10.0	10.4	10.8
	11.2	10.7	10.7	
	(11.5)	(10.35)	(10.55)	
Lab mean	11.3	10.1	9.8	10.4 (overall mean)

a. Perform an analysis of variance for this experiment. Conduct all tests with $\alpha = .05$.

b. Estimate all variance components, and determine their proportional allocation to the total variability.

c. Estimate the average calcium level over all laboratories and assays.

Solution **a.** Using the formulas from Chapter 14, we obtain the sums of squares as follows:

$$\text{TSS} = \sum_{ijk}(y_{ijk} - \bar{y}_{\cdots})^2 = (10.9 - 10.4)^2 + (10.9 - 10.4)^2 + \cdots + (10.7 - 10.4)^2$$
$$= 12.00$$

$$\text{SSA} = \sum_{i} 6(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2 = 6\{(10.3 - 10.4)^2 + (10.1 - 10.4)^2 + (10.8 - 10.4)^2\}$$
$$= 1.56$$

$$\text{SSL} = \sum_j 6(\bar{y}_{.j.} - \bar{y}_{...})^2 = 6\{(11.3 - 10.4)^2 + (10.1 - 10.4)^2 + (9.8 - 10.4)^2\}$$
$$= 7.56$$

$$\text{SSAL} = \sum_{ij} 2(\bar{y}_{ij.} - \bar{y}_{...})^2 - \text{SSA} - \text{SSL} = 2\{(10.9 - 10.4)^2 + (10.15 - 10.4)^2$$
$$+ (9.85 - 10.4)^2 + \cdots + (10.55 - 10.4)^2\} - 1.56 - 7.56 = 1.64$$

$$\text{SSE} = \text{TSS} - \text{SSA} - \text{SSL} - \text{SSAL} = 12.00 - 1.56 - 7.56 - 1.64 = 1.24$$

Our results are summarized in an analysis of variance table in Table 17.15.

TABLE 17.15
AOV table for Example 17.3 experiment

Source	SS	df	MS	EMS
Assay	1.56	2	.78	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2 + 6s_\tau^2$
Lab	7.56	2	3.78	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2 + 6s_\beta^2$
Assay*lab	1.64	4	.41	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2$
Error	1.24	9	.1378	σ_ε^2
Total	12.00	17		

We can proceed with appropriate statistical tests, using the results presented in the AOV table. For the AL interaction, we have

H_0: $\sigma_{\tau\beta}^2 = 0$

H_a: $\sigma_{\tau\beta}^2 > 0$

T.S.: $F = \dfrac{\text{MSAL}}{\text{MSE}} = \dfrac{.41}{.1378} = 2.98$

R.R.: For $\alpha = .05$, we will reject H_0 if F exceeds 3.63, the critical value for F with $\alpha = .05$, $\text{df}_1 = 4$, and $\text{df}_2 = 9$.

Conclusion: There is insufficient evidence to reject H_0, p-value = .08. There does not appear to be a significant interaction between the levels of factors A and L.

For factor L, we have

H_0: $\sigma_\beta^2 = 0$

H_a: $\sigma_\beta^2 > 0$

T.S.: $F = \dfrac{\text{MSL}}{\text{MSAL}} = \dfrac{3.78}{.41} = 9.22$

R.R.: For $\alpha = .05$, we will reject H_0 if F exceeds 6.94, the critical value based on $\alpha = .05$, $\text{df}_1 = 2$, and $\text{df}_2 = 4$.

Conclusion: Because the observed value of F is much larger than 6.94, we reject H_0 and conclude that there is a significant variability in calcium concentrations from lab to lab, p-value = .032.

The test for factor A follows:

H_0: $\sigma_\tau^2 = 0$

H_a: $\sigma_\tau^2 > 0$

T.S.: $F = \dfrac{\text{MSA}}{\text{MSAL}} = \dfrac{.78}{.41} = 1.90$

R.R.: For $\alpha = .05$, we will reject H_0 if F exceeds 6.94, the critical value for $\alpha = .05$, $df_1 = 2$, and $df_2 = 4$.

Conclusion: There is insufficient evidence to indicate a significant variability in calcium determinations from assay to assay, p-value $= .263$.

b. We will next estimate the variance components. Using the MSs and EMSs in Table 17.15, we obtain

$$\hat{\sigma}^2_{\varepsilon} = \text{MSE} = .1378$$

$$\hat{\sigma}^2_{\tau\beta} = (\text{MSAL} - \text{MSE})/n = (.41 - .1378)/2 = .1361$$

$$\hat{\sigma}^2_{\beta} = (\text{MSL} - \text{MSAL})/an = (3.78 - .41)/6 = .5617$$

and

$$\hat{\sigma}^2_{\tau} = (\text{MSA} - \text{MSAL})/bn = (.78 - .41)/6 = .0617$$

Also, from the random-effects model for two factors having randomly selected levels, we have

$$E(y_{ijk}) = \mu \text{ and } \sigma^2_y = \sigma^2_{\tau} + \sigma^2_{\beta} + \sigma^2_{\tau b} + \sigma^2_{\varepsilon}$$

Thus, we have

$$\sigma^2_y = .0617 + .5617 + .1361 + .1378 = .8973$$

We can then proportionally allocate the total variability $\hat{\sigma}^2_y$ into the four sources of variability: assays, laboratories, the interaction, and experimental error, shown in Table 17.16.

TABLE 17.16
Proportional allocation of total variance

Source of Variance	Estimator	Proportion of Total
Assays	.0617	.0617/.8973 = .069
Labs	.5617	.5617/.8973 = .626
Interaction	.1361	.1361/.8973 = .152
Error	.1378	.1378/.8973 = .154
Totals	.8973	1.0

c. Because there was a significant variability in the determinations of calcium in the samples, the estimation of an overall mean level μ would not be of interest to the researchers. However, to illustrate the methodology, we will proceed with this example. The point estimator of μ and its estimated standard error are given by

$$\hat{\mu} = \bar{y}_{\ldots} = 10.4 \text{ and } \text{SE}(\hat{\mu}) = \sqrt{(\text{MSA} + \text{MSL} - \text{MSAL})/abn} = .4802$$

We can then construct a $100(1 - \alpha)\%$ confidence interval for μ, as given here.

$$\bar{y}_{\ldots} \pm t_{\alpha/2,\, \text{df}_{\text{Approx.}}} \sqrt{(\text{MSA} + \text{MSL} - \text{MSAL})/abn} = 10.4 \pm (t_{.025,\, \text{df}_{\text{Approx.}}})(.4802)$$

where the degrees of freedom for the t tables are obtained from the Satterthwaite approximation:

$$\text{df}_{\text{Approx.}} = \frac{(\text{MSA} + \text{MSL} - \text{MSAL})^2}{(\text{MSA})^2/(a-1) + (\text{MSL})^2/(b-1) + (\text{MSAL})^2/(a-1)(b-1)}$$

$$= \frac{(4.15)^2}{(.78)^2/2 + (3.78)^2/2 + (.41)^2/4} = 2.3$$

We take the largest integer less than or equal to $df_{Approx.}$; thus, $df_{Approx.} = 2$. Because $t_{.025,2} = 4.303$, the 95% confidence interval for the mean calcium concentration over all assays and laboratories is

$$10.4 \pm (4.303)(.4802)$$

$$10.4 \pm 2.1 = (8.3, 12.5) \blacksquare$$

In this section, we have compared a random-effects model to a fixed-effects model for a completely randomized design and for a completely randomized design with an $a \times b$ factorial treatment structure with n observations per cell. This study has been in no way exhaustive, but it has shown that there are alternatives to a fixed-effects model. A more detailed study of the random-effects model in the following sections will include experiments with factorial treatment structures having more than two factors and the **nested sampling experiment** of Section 17.6. For the latter design, levels of factor B are nested (rather than cross-classified) within levels of factor A. For example, in considering the potency of a chemical, we could sample different manufacturing plants, batches of chemicals within a plant, and determinations within a batch. Note that the factor "batches" is not cross-classified with the factor "plants" because, for example, batch 1 for plant 1 is different from batch 1 for plant 2.

nested sampling experiment

In Section 17.4, we will extend the results of this section to include a mixed model for an $a \times b$ factorial treatment structure with one fixed-effects factor and one random-effects factor.

17.4 Mixed-Effects Models

In Section 17.3, we compared the analysis of variance tables for fixed- and random-effects models for a randomized block design and for a general $a \times b$ factorial treatment structure laid out in a completely randomized design. Suppose, however, that we have a **mixed-effects model** for these same experimental designs, where one effect is fixed and the other is random. For example, in Section 17.3, we considered an experiment to examine the effects of different subjects and different analysts on the DNA content of plaque. If the 10 subjects were selected at random and if the five analysts chosen were the only analysts of interest, we would have a mixed model for a randomized block design with fixed analysts and random subjects.

mixed-effects model

Let us consider a mixed model for a general $a \times b$ factorial treatment structure in a completely randomized design. The model is the same as that given in Section 17.3 except that there are different assumptions:

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$$

where we use the following **conditions** with the levels of factor A fixed and the levels of factor B randomly selected:

conditions

1. μ is the unknown overall mean response.
2. τ_i is a fixed effect corresponding to the ith level of factor A with $\tau_a = 0$.
3. β_j is a random effect due to the jth level of factor B. The β_js have independent normal distributions with mean 0 and variance σ^2_β.
4. $\tau\beta_{ij}$ is a random effect due to the interaction of the ith level of factor A with the jth level of factor B. The $\tau\beta_{ij}$s have independent normal distributions with mean 0 and variance $\sigma^2_{\tau\beta}$.
5. The β_js, $\tau\beta_{ij}$s, and ε_{ijk}s are mutually independent.

TABLE 17.17
AOV table for an $a \times b$ factorial treatment structure, with n observations per cell

Source	SS	df	MS	EMS: Fixed Effects	EMS: Random Effects	EMS: Mixed Effects *A* Fixed, *B* Random
A	SSA	$a-1$	MSA	$\sigma_\varepsilon^2 + bn\theta_\tau$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2 + bn\sigma_\tau^2$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2 + bn\theta_\tau$
B	SSB	$b-1$	MSB	$\sigma_\varepsilon^2 + an\theta_\beta$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2 + an\sigma_\beta^2$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2 + an\sigma_\beta^2$
AB	SSAB	$(a-1)(b-1)$	MSAB	$\sigma_\varepsilon^2 + n\theta_{\tau\beta}$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2$
Error	SSE	$ab(n-1)$	MSE	σ_ε^2	σ_ε^2	σ_ε^2
Totals	TSS	$nab-1$				

Using these assumptions, the analysis of variance table for a fixed, random, or mixed model in a two-factor experiment with replication is as shown in Table 17.17.

The expected mean squares column of Table 17.17 can be helpful in determining appropriate tests of significance. The test for $\sigma_{\tau\beta}^2$ is the same in the mixed-effects model as in the random-effects model.

test for $\sigma_{\tau\beta}^2$

H_0: $\sigma_{\tau\beta}^2 = 0$

H_a: $\sigma_{\tau\beta}^2 > 0$

T.S.: $F = \dfrac{\text{MSAB}}{\text{MSE}}$

R.R.: Based on $\text{df}_1 = (a-1)(b-1)$ and $\text{df}_2 = ab(n-1)$

No matter what the results are of our tests for $\sigma_{\tau\beta}^2$, we can proceed to use the following tests for factors A and B, which follow from entries in the expected mean squares column of Table 17.17. For factor A, we have

test, factor A

H_0: $\tau_1 = \cdots = \tau_a = 0$

H_a: At least one of the τs differs from the rest.

T.S.: $F = \dfrac{\text{MSA}}{\text{MSAB}}$

R.R.: Based on $\text{df}_1 = (a-1)$ and $\text{df}_2 = (a-1)(b-1)$

For factor B, we have

test, factor B

H_0: $\sigma_\beta^2 = 0$

H_a: $\sigma_\beta^2 > 0$

T.S.: $F = \dfrac{\text{MSB}}{\text{MSAB}}$

R.R.: Based on $\text{df}_1 = (b-1)$ and $\text{df}_2 = (a-1)(b-1)$

The analysis of variance procedure outlined for a mixed-effects model for an $a \times b$ factorial treatment structure can be used as well for a randomized block design, where treatments are fixed, blocks are assumed to be random, and there are n observations for each block and treatment. We will illustrate a mixed model in the following example.

EXAMPLE 17.4

A study was designed to evaluate the effectiveness of two different sunscreens (s_1 and s_2) for protecting the skin of persons who want to avoid burning or additional tanning while exposed to the sun. A random sample of 40 subjects (ages 20–25) agreed to participate in the study. For each subject, a 1-inch square was marked off on his or her back, under the shoulder but above the small of the back. Twenty subjects were randomly assigned to each of the two types of sunscreen. A reading based on the color of the skin in the designated square was made prior to the application of a fixed amount of the assigned sunscreen and then again after application and exposure to the sun for a 2-hour period. The company was concerned that the measurement of color is extremely variable and wanted to assess the variability in the readings due to the technician taking the readings. Thus, the company randomly selected 10 technicians from their worldwide staff to participate in the study. Four subjects, two having s_1 and two having s_2, were randomly assigned to each technician for evaluation. The data recorded in Table 17.18 are differences (postexposure minus preexposure) for the subjects in the study. A high response indicates a greater degree of burning.

TABLE 17.18
Data for sunscreen experiment in Example 17.4

	Technician (B)										
Sunscreen (A)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Sun. Mean**
s_1	8.2	3.6	10.7	3.9	12.9	5.5	9.1	13.7	8.1	2.5	7.82
	7.6	3.5	10.3	4.4	12.1	5.9	9.7	13.2	8.7	2.8	
Mean	(7.9)	(3.55)	(10.5)	(4.15)	(12.5)	(5.7)	(9.4)	(13.45)	(8.4)	(2.65)	
s_2	6.1	4.3	9.6	2.3	12.4	4.8	8.3	12.9	8.0	2.1	7.15
	6.8	4.7	9.2	2.5	12.8	4.0	8.6	13.6	7.5	2.5	
Mean	(6.45)	(4.5)	(9.4)	(2.4)	(12.6)	(4.4)	(8.45)	(13.25)	(7.75)	(2.3)	
Tech. Mean	(7.175)	(4.025)	(9.95)	(3.275)	(12.55)	(5.05)	(8.925)	(13.35)	(8.075)	(2.475)	7.485

The experiment is a completely randomized design with two factors, sunscreen type (A), with 2 fixed levels, and technician (B), with 10 randomly selected levels. There are two subjects for each sunscreen–technician combination. Analyze the data to determine any differences in sunscreens and technicians.

Solution We can compute the sums of squares for the sources of variability in the AOV table using the following formulas.

$$\text{TSS} = \sum_{ijk}(y_{ijk} - \bar{y}_{...})^2 = (8.2 - 7.485)^2 + (7.6 - 7.485)^2 + \cdots + (2.5 - 7.485)^2 = 530.59$$

$$\text{SSA} = \sum_{i} 20(\bar{y}_{i..} - \bar{y}_{...})^2 = 20\left\{(7.82 - 7.485)^2 + (7.15 - 7.485)^2\right\} = 4.49$$

$$\text{SSB} = \sum_{j} 4(\bar{y}_{.j.} - \bar{y}_{...})^2 = 4\left\{(7.175 - 7.485)^2 + (4.025 - 7.485)^2 + \cdots + (2.475 - 7.485)^2\right\} = 517.49$$

$$\text{SSAB} = \sum_{ij} 2(\bar{y}_{ij.} - \bar{y}_{...})^2 - \text{SSA} - \text{SSB} = 2\{(7.9 - 7.485)^2 + (3.55 - 7.485)^2 + (10.5 - 7.485)^2 + \cdots + (2.3 - 7.485)^2\} - 4.49 - 517.49 = 5.97$$

$$\text{SSE} = \text{TSS} - \text{SSA} - \text{SSB} - \text{SSAB} = 530.59 - 4.49 - 517.49 - 5.97 = 2.64$$

Substituting $a = 2$, $b = 10$, and $n = 2$ into an AOV table similar to that shown in Table 17.17, we have the results shown in Table 17.19.

TABLE 17.19
AOV table for the data of Example 17.4

Source	SS	df	MS	EMS Mixed Model
A	4.49	1	4.49	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2 + 20\theta_\tau$
B	517.49	9	57.50	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2 + 4\sigma_\beta^2$
AB	5.97	9	.66	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2$
Error	2.64	20	.13	σ_ε^2
Totals	530.59	39		

A test for the random component $\tau\beta_{ij}$ is as follows:

H_0: $\sigma_{\tau\beta}^2 = 0$

H_a: $\sigma_{\tau\beta}^2 > 0$

T.S.: $F = \dfrac{\text{MSAB}}{\text{MSE}} = \dfrac{.66}{.13} = 5.08$

R.R.: For $\alpha = .05$, we will reject H_0 if the computed value of F exceeds 2.39, the value in Appendix Table 8 for $\alpha = .05$, $\text{df}_1 = 9$, and $\text{df}_2 = 20$.

Conclusion: Because 5.08 exceeds 2.39, we reject H_0 and conclude that $\sigma_{\tau\beta}^2 > 0$; that is, there is a significant source of random variation due to the combination of the ith level of A (sunscreens) and the jth level of B (technician), p-value = .0012. We would infer from this that the variations in the determinations of skin color due to technician differences are different for the two types of sunscreen.

We next proceed to evaluate the effects due to the technicians.

H_0: $\sigma_\beta^2 = 0$

H_a: $\sigma_\beta^2 > 0$

T.S.: $F = \dfrac{\text{MSB}}{\text{MSAB}} = \dfrac{57.50}{.66} = 87.12$

R.R.: For $\alpha = .05$, we will reject H_0 if F exceeds 3.18, the value in Appendix Table 8 for $\alpha = .05$, $\text{df}_1 = 9$, and $\text{df}_2 = 9$.

Conclusion: Because 87.12 exceeds 3.18, we reject H_0 and conclude that $\sigma_\beta^2 > 0$. Thus, there is a significant source of random variation due to variability from technician to technician, p-value $< .0001$.

For factor A, we have

H_0: $\tau_1 = \tau_2 = 0$

H_a: $\tau_1 \neq 0$ and/or $\tau_2 = 0$

T.S.: $F = \dfrac{\text{MSA}}{\text{MSAB}} = \dfrac{4.49}{.66} = 6.80$

R.R.: For $\alpha = .05$, we will reject H_0 if F exceeds 5.12, the value in Appendix Table 8 for $\alpha = .05$, $df_1 = 1$, and $df_2 = 9$.

Conclusion: Because $6.80 > 5.12$, we reject H_0 and conclude there is significant evidence that the mean response (postexposure minus preexposure) differs for the two sunscreens. However, as noted previously, there are significant sources of variability due to technicians and the combination of technicians with sunscreens. ■

17.5 Rules for Obtaining Expected Mean Squares

We discussed the AOVs for one- and two-factor experiments for fixed-effects models in Chapter 14 and for random or mixed models earlier in this chapter. We will see in this section that for any k-factor treatment structure of data, with n observations per factor–level combination, it is possible to write expected mean squares for all main effects and interactions for fixed, random, or mixed models using some rather simple rules. *The importance of these rules is that, having written down the expected mean squares for an unfamiliar experimental design, we often can construct appropriate F tests.* The assumptions for the fixed and random models will be the same as we have used in describing fixed, random, and mixed models in previous sections.

classifying interactions

Two rules for **classifying interactions** as fixed or random effects are needed before we can proceed with the rules for obtaining expected mean squares.

Rules for the Classification of Interactions

1. If a fixed effect interacts with another fixed effect, the resulting interaction term is a fixed effect.
2. If a random effect interacts with another effect (fixed or random), the resulting interaction term is a random component.

EXAMPLE 17.5

Consider an experiment with two factors, A with four levels and B with six levels. Suppose we have a completely randomized design with four replications for each of the $t = 24$ treatments. For each of the following situations, classify the AB interaction as fixed or random:

1. The levels of A and B are the only levels of interest to the researcher.
2. The four levels of factor A are the only levels of interest to the researcher, but the six levels of factor B are randomly selected from a population of levels.
3. The four levels of factor A are randomly selected from a population of levels, and the six levels of factor B are randomly selected from a population of levels.

Solution First, we need to determine if the levels of factors A and B are fixed or random; then we apply the rules for the classification of interactions to reach the following conclusions:

1. Both factors A and B have fixed effects; therefore, their interaction, AB, has fixed effects.
2. Factor A has fixed effects and factor B has random effects; therefore, their interaction, AB, has random effects.
3. Both factor A and factor B have random effects; therefore, their interaction, AB, has random effects. ■

EXAMPLE 17.6

Consider an experiment with three factors, A with six levels, B with four levels, and C with three levels. Suppose we have a completely randomized design with two replications for each of the $t = 72$ treatments. The six levels of factor A are randomly selected from a population of levels, whereas the four levels of factor B and the three levels of factor C are the only levels of interest to the researcher. Classify the two-way interactions AB, AC, and BC and three-way interaction ABC as fixed or random.

Solution We apply the classification rules with factor A having random effects and factors B and C having fixed effects.

- A has random effects and B has fixed effects; therefore, AB has random effects.
- A has random effects and C has fixed effects; therefore, AC has random effects.
- B has fixed effects and C has fixed effects; therefore, BC has fixed effects.
- A has random effects and B and C have fixed effects; therefore, ABC has random effects. ■

The rules for obtaining the expected mean squares will be given next. These rules apply to most balanced designs with equal numbers of replications per treatment. The number of levels of each factor must remain constant within the balanced design. The rules are applicable to factorial treatment structures, nested treatment structures, and mixtures of factorial and nested treatment structures. These rules are consistent with the expected mean squares that can be obtained from most statistics software programs (e.g., SAS, R, and Minitab). The rules will be illustrated using a two-factor experiment with n replications, factor A having a randomly selected levels and factor B having b fixed levels.

Rules for Obtaining Expected Mean Squares

1. Write the model for a completely randomized design with an $a \times b$ factorial treatment structure where factor A has random levels and factor B has fixed levels. The model is

$$y_{ijk} = \mu + t_i + \beta_j + \tau\beta_{ij} + \varepsilon_{k[ij]}$$

Note: We use brackets in the ε-term to indicate that there are $k = 1, \ldots, n$ unique experimental units for each of the factor–level combinations of factors A and B (i.e., for each selection of (i, j)).

2. Construct a two-way table consisting of
   a. A row for each term in the model, excluding μ, including the term from the model and the corresponding source of variation from the AOV table, and
   b. A column for each subscript included in the model.
3. Over each column subscript, write the number of factor levels associated with the subscript, and place either an "R" if the factor levels are random or an "F" if the factor levels are fixed.
4. Add another column with entries for the appropriate fixed variance component (θ) or random variance component (σ) for the source of variation represented by that row in the table. The following table, where factor A is random and factor B is fixed, illustrates these rules:

Source		R *a* *i*	F *b* *j*	R *n* *k*	Component
A	τ_i				σ^2_τ
B	β_j				θ_β
AB	$\tau\beta_{ij}$				$\sigma^2_{\tau\beta}$
Error	$\varepsilon_{k[ij]}$				σ^2_ε

5. For each row, if the column subscript does not appear in the effect labeling the row, enter the number of levels corresponding to the subscript heading the column. Otherwise, leave the space blank.

Source		R *a* *i*	F *b* *j*	R *n* *k*	Component
A	τ_i		b	n	σ^2_τ
B	β_j	a		n	θ_β
AB	$\tau\beta_{ij}$			n	$\sigma^2_{\tau\beta}$
Error	$\varepsilon_{k[ij]}$				σ^2_ε

6. For rows having an effect containing brackets in the subscript, place a 1 under the column(s) with a subscript included inside the brackets.

Source		R *a* *i*	F *b* *j*	R *n* *k*	Component
A	τ_i		b	n	σ^2_τ
B	β_j	a		n	θ_β
AB	$\tau\beta_{ij}$			n	$\sigma^2_{\tau\beta}$
Error	$\varepsilon_{k[ij]}$	1	1		σ^2_ε

7. a. For each row in which the component of variance is a *fixed* component, a θ term, enter a 0 in the column headed by an F and having a subscript matching the row subscript.
   b. Enter a 1 in all remaining cells.

TABLE 17.20
Expected mean square table

Source		R *a* *i*	F *b* *j*	R *n* *k*	Component
A	τ_i	1	b	n	σ_τ^2
B	β_j	a	0	n	θ_β
AB	$\tau\beta_{ij}$	1	1	n	$\sigma_{\tau\beta}^2$
Error	$\varepsilon_{k[ij]}$	1	1	1	σ_ε^2

8. To obtain the expected mean square for a specified source of variation (we will illustrate using E(MSA):

a. Include σ_ε^2 with a coefficient of 1 in all expected mean squares.

b. Include in the expected mean square only those variance components whose corresponding model terms include the subscripts of the effect under consideration.

- For E(MSA), the effect is τ_i; hence, include the components σ_τ^2 and $\sigma_{\tau\beta}^2$ associated with τ_i and $\tau\beta_{ij}$, respectively, because they both have an i in their subscripts. Remember to also include σ_ε^2.

c. Cover the columns containing nonbracketed subscripts for the effect under consideration.

- For τ_i, cover the column headed by i; for β_j, cover the column headed by j; for $\tau\beta_{ij}$, cover the columns headed by both i and j; and for $\varepsilon_{k[ij]}$, cover the column headed by k.

d. The coefficient for each component in the expected mean square is the product of the uncovered columns of the row for the effect under consideration.

- For E(MSA), the effect is τ_i, so the column with i is covered. Therefore, the coefficient for σ_τ^2 is obtained by multiplying the entries in the columns headed by j and k—that is, $b \times n$—and the coefficient for $\sigma_{\tau\beta}^2$ is $1 \times n$. Thus,

$$E(\text{MSA}) = \sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2 + bn\sigma_\tau^2$$

EXAMPLE 17.7

Compute E(MSB) and E(MSAB) for a two-factor experiment with a randomly selected levels of A, b fixed levels of B, and n observations per factor–level combination.

Solution Refer to the expected mean squares rules just given and Table 17.20. For E(MSB):

a. Include σ_ε^2 with a coefficient of 1.

b. Include in the expected mean square only those variance components whose corresponding model terms include the subscripts of the effect under consideration.

- For E(MSB), the effect is β_j; hence, include the components θ_β and $\sigma_{\tau\beta}^2$ associated with β and $\tau\beta_{ij}$, respectively, because they both have a j in their subscripts.

c. Cover the columns containing nonbracketed subscripts for the effect under consideration.

- For β_j, cover the column headed by j.

d. The coefficient for each component in the expected mean square is the product of the uncovered columns of the row for the effect under consideration.

- For $E(\text{MSB})$, the effect is β_j, so the column with j is covered. Therefore, the coefficient for θ_β is obtained by multiplying the entries in the columns headed by i and k—that is, $a \times n$—and the coefficient for $\sigma^2_{\tau\beta}$ is $1 \times n$. Thus,

$$E(\text{MSB}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + an\theta_\beta$$

$$\text{where } \theta_\beta = \frac{1}{b-1}\sum_{j=1}^{b}(\mu_{\cdot j} - \mu_{\cdot\cdot})^2$$

For $E(\text{MSAB})$:

a. Include σ^2_ε with a coefficient of 1.

b. Include in the expected mean square only those variance components whose corresponding model terms include the subscripts of the effect under consideration.

- For $E(\text{MSAB})$, the effect is $\tau\beta_{ij}$; hence, include just the component $\sigma^2_{\tau\beta}$ associated with $\tau\beta_{ij}$.

c. Cover the columns containing nonbracketed subscripts for the effect under consideration.

- For $\tau\beta_{ij}$, cover the columns headed by i and j.

d. The coefficient for each component in the expected mean square is the product of the uncovered columns of the row for the effect under consideration.

- For $E(\text{MSAB})$, the effect is $\tau\beta_{ij}$, so cover the columns headed by i and j and obtain the coefficient for $\sigma^2_{\tau\beta}$ as n. Thus,

$$E(\text{MSAB}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta}$$ ■

EXAMPLE 17.8

Obtain the expected mean squares for a factorial treatment structure with a fixed levels of factor A, b randomly selected levels of factor B, and n observations per factor–level combination.

Solution We need to obtain $E(\text{MSA})$, $E(\text{MSB})$, $E(\text{MSAB})$, and $E(\text{MSE})$. The expected mean square table is shown in Table 17.21.

TABLE 17.21
Expected mean square table for Example 17.8

Source		F a i	R b j	R n k	Component
A	τ_i	0	b	n	θ_τ
B	β_j	a	1	n	σ^2_β
AB	$\tau\beta_{ij}$	1	1	n	$\sigma^2_{\tau\beta}$
Error	$\varepsilon_{k[ij]}$	1	1	1	σ^2_ε

For $E(\text{MSA})$:

a. Include σ^2_ε with a coefficient of 1.

b. For $E(\text{MSA})$, the effect is τ_i; hence, include the components θ_τ and $\sigma^2_{\tau b}$.

c. For τ_i, cover the column headed by i.
d. The coefficient for θ_τ is obtained by multiplying the entries in the columns headed by j and k—that is, $b \times n$—and the coefficient for $\sigma^2_{\tau\beta}$ is $1 \times n$. Thus,

$$E(\text{MSA}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + bn\theta_\tau$$

$$\text{where } \theta_\tau = \frac{1}{a-1}\sum_{i=1}^{a}(\mu_{i.} - \mu_{..})^2$$

For E(MSB):

a. Include σ^2_ε with a coefficient of 1.
b. For E(MSB), the effect is β_j; hence, include the components σ^2_β and $\sigma^2_{\tau\beta}$ associated with β_j and $\tau\beta_{ij}$, respectively.
c. For β_j, cover the column headed by j.
d. The coefficient for σ^2_β is obtained by multiplying the entries in the columns headed by i and k—that is, $a \times n$—and the coefficient for $\sigma^2_{\tau\beta}$ is $1 \times n$. Thus,

$$E(\text{MSB}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + an\sigma^2_\beta$$

For E(MSAB):

a. Include σ^2_ε with a coefficient of 1.
b. For E(MSAB), the effect is $\tau\beta_{ij}$; hence, include the component $\sigma^2_{\tau\beta}$ associated with $\tau\beta_{ij}$.
c. For $\tau\beta_{ij}$, cover the columns headed by i and j.
d. The coefficient for $\sigma^2_{\tau\beta}$ is n. Thus,

$$E(\text{MSAB}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta}$$

For E(MSE):

a. Include σ^2_ε with a coefficient of 1.
b. For E(MSE), the effect is $\varepsilon_{k[ij]}$; hence, include the component σ^2_ε.
c. For $\varepsilon_{k[ij]}$, cover the column headed by k.
d. For E(MSE), the effect is $\varepsilon_{k[ij]}$, so cover the column headed by k, and obtain the coefficient for σ^2_ε as 1×1. Thus,

$$E(\text{MSE}) = \sigma^2_\varepsilon$$

Tables 17.22, 17.23 and 17.24 provide the expected mean squares for three arrangements of a two-factor experiment.

TABLE 17.22
AOV table with expected mean squares for factor A random and factor B random

Source	df	Expected Mean Square
A	$a-1$	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + bn\sigma^2_\tau$
B	$b-1$	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + an\sigma^2_\beta$
AB	$(a-1)(b-1)$	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta}$
Error	$(n-1)ab$	σ^2_ε
Total	$nab-1$	

TABLE 17.23
AOV table with expected mean squares for factor A fixed and factor B random

Source	df	Expected Mean Square
A	$a - 1$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2 + bn\theta_\tau$
B	$b - 1$	$\sigma_\varepsilon^2 + n\sigma_{\tau b}^2 + an\sigma_\beta^2$
AB	$(a - 1)(b - 1)$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2$
Error	$(n - 1)ab$	σ_ε^2
Total	$nab - 1$	

TABLE 17.24
AOV table with expected mean squares for factor A random and factor B fixed

Source	df	Expected Mean Square
A	$a - 1$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2 + bn\sigma_\tau^2$
B	$b - 1$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2 + an\theta_\beta$
AB	$(a - 1)(b - 1)$	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2$
Error	$(n - 1)ab$	σ_ε^2
Total	$nab - 1$	

Previously, we have been concerned with only fixed-effects models. For these models, the test statistics are always formed using the affected mean square in the numerator divided by MSE. However, for random and mixed models, the test statistics do not all have MSE in the numerator. The test statistic for interaction is $F = \text{MSAB/MSE}$, which is the same for the fixed, random, and mixed models. The test for the main effect of factor A is $F = \text{MSA/MSAB}$, and the test statistic for the main effect of factor B is $F = \text{MSB/MSAB}$ for all cases except when both factor A and factor B are fixed. These results are obtained by placing in the denominator the mean square having the same expected mean square as the expression for the affected mean square obtained under the null hypothesis. For example, consider the case with factor A fixed and factor B random, as displayed in Table 17.23. To test for a main effect of factor A, H_0: $\theta_\tau = 0$ versus H_a: $\theta_\tau \neq 0$, we determine from Table 17.23 that $E(\text{MSA}) = \sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2$ under H_0; that is, we set $\theta_\tau = 0$. This is the same as the expression for $E(\text{MSAB})$; therefore, the test statistic is $F = \text{MSA/MSAB}$. Similarly, to test for a main effect of factor B, H_0: $\sigma_\beta^2 = 0$ versus H_a: $\sigma_\beta^2 \neq 0$, we determine from Table 17.23 that $E(\text{MSB}) = \sigma_\varepsilon^2 + n\sigma_{\tau\beta}^2$ under H_0. This is the same as the expression for $E(\text{MSAB})$; therefore, the test statistic is $F = \text{MSB/MSAB}$.

The same rules used for the factorial treatment structure with two factors can also be used for more-complicated experiments, and although the rules may seem a bit cumbersome, with practice they are quite easy to use. We will give two more examples using a factorial treatment structure with three factors. For additional details regarding assumptions, derivations, and more-complicated applications, see Kuehl (2000).

EXAMPLE 17.9

Provide the expected mean squares for a $6 \times 5 \times 4$ factorial treatment structure with $n = 3$ observations per factor–level combination. In the experiment, factors A and B have fixed levels, but factor C has randomly selected levels.

Solution The model for this experiment is given here along with the corresponding expected mean squares for each of the sources of variation:

$$y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_k + \tau\beta_{ij} + \tau\gamma_{ik} + \beta\gamma_{jk} + \tau\beta\gamma_{ijk} + \varepsilon_{l[ijk]}$$

The expected mean square are obtained from Table 17.25.

TABLE 17.25
Expected mean squares table for factors A and B fixed, factor C random

Source		F a i	F b j	R c k	R n l	Component
A	τ_i	0	b	c	n	θ_τ
B	β_j	a	0	c	n	θ_β
C	γ_k	a	b	1	n	σ^2_γ
AB	$\tau\beta_{ij}$	0	0	c	n	$\theta_{\tau\beta}$
AC	$\tau\gamma_{ik}$	1	b	1	n	$\sigma^2_{\tau\gamma}$
BC	$\beta\gamma_{jk}$	a	1	1	n	$\sigma^2_{\beta\gamma}$
ABC	$\tau\beta\gamma_{ijk}$	1	1	1	n	$\sigma^2_{\tau\beta\gamma}$
Error	$\varepsilon_{l[ijk]}$	1	1	1	1	σ^2_ε

For E(MSA):

a. Include σ^2_ε with a coefficient of 1.
b. For E(MSA), the effect is τ_i; hence, include the components θ_τ, $\theta_{\tau\beta}$, $\sigma^2_{\tau\gamma}$, and $\sigma^2_{\tau\beta\gamma}$.
c. For τ_i, cover the column headed by i.
d. The coefficient for each component is obtained by multiplying the entries in the columns headed by j, k, and l—that is, $b \times c \times n$: The coefficient for $\theta_{\tau\beta}$ is $0 \times c \times n$; the coefficient for $\sigma^2_{\tau\gamma}$ is $b \times 1 \times n$; and the coefficient for $\sigma^2_{\tau\beta\gamma}$ is $1 \times 1 \times n$. Thus,

$$E(\text{MSA}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta\gamma} + bn\sigma^2_{\tau\gamma} + bcn\theta_\tau$$

For E(MSB):

a. Include σ^2_ε with a coefficient of 1.
b. For E(MSB), the effect is β_j; hence, include the components, θ_β, $\theta_{\tau\beta}$, $\sigma^2_{\beta\gamma}$, and $\sigma^2_{\tau\beta\gamma}$.
c. For β_j, cover the column headed by j.
d. The coefficient for each component is obtained by multiplying the entries in the columns headed by i, k and l: The coefficient for θ_β is $a \times c \times n$; the coefficient for $\theta_{\tau\beta}$ is $0 \times c \times n$; the coefficient for $\sigma^2_{\beta\gamma}$ is $a \times 1 \times n$; and the coefficient for $\sigma^2_{\tau\beta\gamma}$ is $1 \times 1 \times n$. Thus,

$$E(\text{MSB}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta\gamma} + an\sigma^2_{\beta\gamma} + acn\theta_\beta$$

For E(MSC):

a. Include σ^2_ε with a coefficient of 1.
b. For E(MSC), the effect is γ_k; hence, include the components σ^2_γ, $\sigma^2_{\tau\gamma}$, $\sigma^2_{\beta\gamma}$, and $\sigma^2_{\tau\beta\gamma}$.
c. For γ_k, cover the column headed by k.

d. The coefficient for each component is obtained by multiplying the entries in the columns headed by i, j and l: The coefficient for σ^2_γ is $a \times b \times n$; the coefficient for $\sigma^2_{\tau\gamma}$ is $1 \times b \times n$; the coefficient for $\sigma^2_{\beta\gamma}$ is $a \times 1 \times n$; and the coefficient for $\sigma^2_{\tau\beta\gamma}$ is $1 \times 1 \times n$. Thus,

$$E(\text{MSC}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta\gamma} + an\sigma^2_{\beta\gamma} + bn\sigma^2_{\tau\gamma} + abn\sigma^2_\gamma$$

For E(MSAB):

a. Include σ^2_ε with a coefficient of 1.
b. For E(MSAB), the effect is $\tau\beta_{ij}$; hence, include the components $\theta_{\tau\beta}$ and $\sigma^2_{\tau\beta\gamma}$.
c. For $\tau\beta_{ij}$, cover the columns headed by i and j.
d. The coefficient for $\theta_{\tau\beta}$ is $c \times n$, and the coefficient for $\sigma^2_{\tau\beta\gamma}$ is $1 \times n$. Thus,

$$E(\text{MSAB}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta\gamma} + cn\theta^2_{\tau\beta}$$

In a similar fashion, we obtain

$$E(\text{MSAC}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta\gamma} + bn\sigma^2_{\tau\gamma}$$
$$E(\text{MSBC}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta\gamma} + an\sigma^2_{\beta\gamma}$$
$$E(\text{MSABC}) = \sigma^2_\varepsilon + n\sigma^2_{\tau\beta\gamma}$$
$$E(\text{MSE}) = \sigma^2_\varepsilon$$

A summary of the expected mean squares, which we have computed using the EMS rules, for the $6 \times 5 \times 4$ factorial experiment with $a = 6$, $b = 5$, $c = 4$, and $n = 3$ and with factors A and B fixed but factor C random is shown in Table 17.26. We have included the denominator of the valid F test for testing whether this source of variation is significant. An * indicates a variance component for which there is not a valid F test.

TABLE 17.26
Partial AOV for Example 17.9. Factors A and B fixed, factor C random

Source	EMS	Denominator of F
A	$\sigma^2_\varepsilon + 3\sigma^2_{\tau\beta\gamma} + 15\sigma^2_{\tau\gamma} + 60\theta_\tau$	MSAC
B	$\sigma^2_\varepsilon + 3\sigma^2_{\tau\beta\gamma} + 18\sigma^2_{\beta\gamma} + 72\theta_\beta$	MSBC
C	$\sigma^2_\varepsilon + 3\sigma^2_{\tau\beta\gamma} + 15\sigma^2_{\tau\gamma} + 18\sigma^2_{\beta\gamma} + 90\sigma^2_\gamma$	*
AB	$\sigma^2_\varepsilon + 3\sigma^2_{\tau\beta\gamma} + 12\theta_{\tau\beta}$	MSABC
AC	$\sigma^2_\varepsilon + 3\sigma^2_{\tau\beta\gamma} + 15\sigma^2_{\tau\gamma}$	MSABC
BC	$\sigma^2_\varepsilon + 3\sigma^2_{\tau\beta\gamma} + 18\sigma^2_{\beta\gamma}$	MSABC
ABC	$\sigma^2_\varepsilon + 3\sigma^2_{\tau\beta\gamma}$	MSE
Error	σ^2_ε	*

■

EXAMPLE 17.10

Refer to Example 17.9. Find an appropriate F test statistic for testing each of the following:

a. Main effect of factor A
b. Main effect of factor C
c. Interaction of factors A and C

Solution Using the expected mean squares listed in Table 17.26, we can find the following test statistics.

a. The test for a main effect of factor A has null hypothesis H_0: $\theta_\tau = 0$. Under H_0, $E(\text{MSA}) = \sigma_\varepsilon^2 + 3\sigma_{\tau\beta c}^2 + 15\sigma_{\tau c}^2$, which is the same expression as $E(\text{MSAC})$. Therefore, the test statistic is $F = \text{MSA}/\text{MSAC}$ with df $= a - 1, (a - 1)(c - 1)$.

b. The test for a main effect of factor C has null hypothesis H_0: $\sigma_\gamma = 0$. Under H_0, $E(\text{MSC}) = \sigma_\varepsilon^2 + 3\sigma_{\tau\beta\gamma}^2 + 15\sigma_{\tau\gamma}^2 + 18\sigma_{\beta\gamma}^2$. There is no other source of variation that has this expression as its expected mean square. Therefore, there is no exact F test available. There are several approximate F tests available in this situation (see Kuehl, 2000).

c. The test for an interaction between factors A and C has null hypothesis H_0: $\sigma_{\tau\gamma} = 0$. Under H_0, $E(\text{MSAC}) = \sigma_\varepsilon^2 + \sigma_{\tau\beta\gamma}^2$, which is the same expression as $E(\text{MSABC})$. Therefore, the test statistic is $F = \text{MSAC}/\text{MSABC}$ with df $= (a - 1)(c - 1), (a - 1)\,(b - 1)(c - 1)$. ■

We can always obtain valid tests for all sources of variability in fixed-effects models, but this is not true for some random-effects and mixed-effects models, as was demonstrated in Example 17.10. Tables 17.27, 17.28, 17.29, and 17.30 display the EMS for several three-factor experiments. In these tables, we provide the denominator of the F test for those variance components having valid F tests. An * indicates a variance component for which there is not a valid F test. Approximate F tests

TABLE 17.27
Three-factor $a \times b \times c$ design with all factors fixed and n replications

All Factors Fixed		
Source	**EMS**	**Denominator of *F***
A	$\sigma_\varepsilon^2 + bcn\theta_\tau$	MSE
B	$\sigma_\varepsilon^2 + acn\theta_\beta$	MSE
C	$\sigma_\varepsilon^2 + abn\theta_\tau$	MSE
AB	$\sigma_\varepsilon^2 + cn\theta_{\tau\beta}$	MSE
AC	$\sigma_\varepsilon^2 + bn\theta_{\tau\gamma}$	MSE
BC	$\sigma_\varepsilon^2 + an\theta_{\beta\gamma}$	MSE
ABC	$\sigma_\varepsilon^2 + n\theta_{\tau\beta\gamma}$	MSE
Error	σ_ε^2	*

TABLE 17.28
Three-factor $a \times b \times c$ design with all factors random and n replications

All Factors Random		
Source	**EMS**	**Denominator of *F***
A	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + cn\sigma_{\tau\beta}^2 + bn\sigma_{\tau\gamma}^2 + bcn\sigma_\tau^2$	*
B	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + cn\sigma_{\tau\beta}^2 + an\sigma_{\beta\gamma}^2 + acn\sigma_\beta^2$	*
C	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + bn\sigma_{\tau\gamma}^2 + an\sigma_{\beta\gamma}^2 + abn\sigma_\gamma^2$	*
AB	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + cn\sigma_{\tau\beta}^2$	MSABC
AC	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + bn\sigma_{\tau\gamma}^2$	MSABC
BC	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + an\sigma_{\beta\gamma}^2$	MSABC
ABC	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2$	MSE
Error	σ_ε^2	*

TABLE 17.29
Three-factor $a \times b \times c$ design with factors A and B random, factor C fixed, and n replications

Factors A and B Random, Factor C Fixed		
Source	**EMS**	**Denominator of *F***
A	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + cn\sigma_{\tau\beta}^2 + bn\sigma_{\tau\gamma}^2 + bcn\sigma_\tau^2$	*
B	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + cn\sigma_{\tau\beta}^2 + an\sigma_{\beta\gamma}^2 + acn\sigma_\beta^2$	*
C	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + bn\sigma_{\tau\gamma}^2 + an\sigma_{\beta\gamma}^2 + abn\theta_\gamma$	*
AB	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + cn\sigma_{\tau\beta}^2$	MSABC
AC	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + bn\sigma_{\tau\gamma}^2$	MSABC
BC	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + an\sigma_{\beta\gamma}^2$	MSABC
ABC	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2$	MSE
Error	σ_ε^2	*

TABLE 17.30
Three-factor $a \times b \times c$ design with factor A random, factors B and C fixed, and n replications

Factor A Random, Factors B and C Fixed		
Source	**EMS**	**Denominator of *F***
A	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + cn\sigma_{\tau\beta}^2 + bn\sigma_{\tau\gamma}^2 + bcn\sigma_\tau^2$	*
B	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + cn\sigma_{\tau\beta}^2 + acn\theta_\beta$	MSAB
C	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + bn\sigma_{\tau\gamma}^2 + abn\theta_\gamma$	MSAC
AB	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + cn\sigma_{\tau\beta}^2$	MSABC
AC	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + bn\sigma_{\tau\gamma}^2$	MSABC
BC	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2 + an\theta_{\beta\gamma}$	MSABC
ABC	$\sigma_\varepsilon^2 + n\sigma_{\tau\beta\gamma}^2$	MSE
Error	σ_ε^2	*

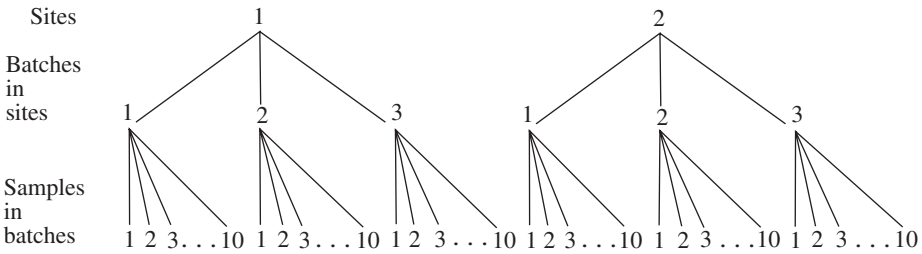
can be constructed for sources of variability in random-effects and mixedeffects models where no valid F test is available. These tests are available in some of the computer software programs—for example, SAS and R. A discussion of these tests can be found in Kuehl (2000).

The estimation of variance components was illustrated in Sections 17.2 and 17.3. Mean squares can be equated to expected mean squares in order to obtain estimates of variance components in random-effects and mixed-effects models for balanced designs following the procedure that we introduced in these earlier sections. Many computer software programs—for example, SAS and R—will carry out these calculations. The problem of estimating variance components for unbalanced designs is a complex one and is beyond the scope of this text. A detailed discussion of this topic can be found in Searle, Casella, and McCulloch (1992).

17.6 Nested Factors

Sometimes in an experiment one factor is "nested" within another. This can be illustrated with the following example. A pharmaceutical company conducted tests to determine the stability of its product (under room-temperature conditions) at a specific point in time. Two manufacturing sites were used. At each site, a random sample of three batches of the product was obtained, and additional random samples of 10 different tablets were obtained from each batch. The design can be represented as shown in Figure 17.1.

FIGURE 17.1
Two-factor experiment with batches nested in sites



Although this might look like the usual two-factor experiment with sites (factor A) and batches (factor B), note that the three batches taken from site 1 are different from the three batches taken from site 2. In this sense, factor B (batches) is said to be *nested* in factor A (sites). In order to distinguish between experiments involving crossed factors and nested factors, consider the following definitions.

DEFINITION 17.4

In an $a \times b \times c$ factorial experiment, the factors A, B, and C are said to be **crossed** if the physical properties of the b levels of factor B are identical for all levels of factor A and the c levels of factor C are identical for all levels of factor B. We denote crossed factors by A × B × C.

This would not be true in the pharmaceutical example described previously. Designate factor A to be the two sites, factor B to be the three batches at each site, and factor C to be the 10 tables from each batch at each site. The three batches at site 1 are potentially not the same as the three batches at site 2. Likewise, the 10 tablets from batch 1 at site 1 are potentially quite different from the 10 tablets from batch 1 at site 2. The levels of factor B, batches, are dependent on which site they came from, and the levels of factor C, tablets, are dependent on which batch they came from and which site they came from. Thus, we have the following definition.

DEFINITION 17.5

In an experiment involving the factors A, B, and C, factor B is said to be **nested** within the levels of factor A if the physical properties of the b levels of factor B vary depending on which level of factor A it is associated with; factor C is said to be **nested** within the levels of factors A and B if the physical properties of the c levels of factor C vary depending on which level of factor A and which level of factor B it is associated with. We denote nested factors as B(A) for factor B nested within factor A and C(A, B) for factor C nested within factors A and B.

In the pharmaceutical example, the batches are nested within the sites; that is, factor B is nested within factor A. Also, the tablets, factor C, are nested within the levels of factor B and hence the levels of factor A. That is, the three batches within a site are unique to that site, and the 10 tablets within a batch are unique to that batch and hence also unique to the site associated with that batch.

For an experimental situation having factor B nested within factor A, it will be impossible to evaluate the effect of the interaction of factor B with factor A because each level of factor B does not appear with each level of factor A, as it would in a factorial (crossed) arrangement of factors A and B.

TABLE 17.31
AOV table for a two-factor experiment (n observations per cell) with factor B nested within factor A

Source	SS	df	MS	Expected Mean Square: A&B Fixed	Expected Mean Square: A Fixed, B Random	Expected Mean Square: A&B Random
A	SSA	$a-1$	MSA	$\sigma_\varepsilon^2 + bn\theta_\tau$	$\sigma_\varepsilon^2 + n\sigma_{\beta(\tau)}^2 + bn\theta_\tau$	$\sigma_\varepsilon^2 + n\sigma_{\beta(\tau)}^2 + bn\sigma_\tau^2$
B(A)	SSB(A)	$a(b-1)$	MSB(A)	$\sigma_\varepsilon^2 + n\theta_{\beta(\tau)}$	$\sigma_\varepsilon^2 + n\sigma_{\beta(\tau)}^2$	$\sigma_\varepsilon^2 + n\sigma_{\beta(\tau)}^2$
Error	SSE	$ab(n-1)$	MSE	σ_ε^2	σ_ε^2	σ_ε^2

The general model for a two-factor experiment (n observations per cell) where factor B is nested in factor A can be written as

$$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \varepsilon_{ijk} \qquad \begin{aligned} i &= 1, 2, \ldots, a \\ j &= 1, 2, \ldots, b \\ k &= 1, 2, \ldots, n \end{aligned}$$

Note that this model is similar to the model for the two-factor experiment of Section 17.3 except that there is no interaction term $\tau\beta_{ij}$ and the term for factor B, $\beta_{j(i)}$, is subscripted to denote the jth level of factor B is nested in the ith level of factor A. The analysis of variance table for this design is shown in Table 17.31.

The sums of squares in the AOV table are computed using the formulas given here.

$$\text{TSS} = \sum_{ijk} (y_{ijk} - \bar{y}_{...})^2$$

$$\text{SSA} = \sum_{i} bn(\bar{y}_{i..} - \bar{y}_{...})^2$$

$$\text{SSB(A)} = \sum_{i}\sum_{j} n(\bar{y}_{ij.} - \bar{y}_{i..})^2$$

$$\text{SSE} = \text{TSS} - \text{SSA} - \text{SSB(A)}$$

Three of the more common situations are shown in Table 17.31 with the expected mean squares. Note the following in particular:

1. The F test for factor B(A), H_0: $\theta_{\beta(\tau)} = 0$ or H_0: $\sigma_{\beta(\tau)}^2 = 0$, is always
$$F = \frac{\text{MSB(A)}}{\text{MSE}}$$
2. The F test for factor A in the fixed-effects model, H_0: $\theta_\tau = 0$, is
$$F = \frac{\text{MSA}}{\text{MSE}}$$
For the random- and mixed-effects models, however, the corresponding test for factor A, H_0: $\sigma_\tau^2 = 0$ or H_0: $\theta_\tau = 0$, is
$$F = \frac{\text{MSA}}{\text{MSB(A)}}$$
3. When $n = 1$, there is no test for factor B(A), but we can test for factor A in the random- and mixed-effects models using
$$F = \frac{\text{MSA}}{\text{MSB(A)}}$$

EXAMPLE 17.11

Researchers conducted an experiment to determine the content uniformity of film-coated tablets produced for a cardiovascular drug used to lower blood pressure. They obtained a random sample of three batches from each of two blending sites; within each batch, they assayed a random sample of five tablets to determine content uniformity. The data are shown here:

Site	1			2		
Batches within each site	1	2	3	1	2	3
Tablets within each batch	5.03	4.64	5.10	5.05	5.46	4.90
	5.10	4.73	5.15	4.96	5.15	4.95
	5.25	4.82	5.20	5.12	5.18	4.86
	4.98	4.95	5.08	5.12	5.18	4.86
	5.05	5.06	5.14	5.05	5.11	5.07

a. Run an analysis of variance. Use $\alpha = .05$.
b. Is there evidence to indicate batch-to-batch variability in content uniformity? Does the F test run depend on whether we assume batches are fixed or random?
c. Draw conclusions about batch.

Solution

a. For these data, we have $a = 2$ blending sites, $b = 3$ batches within each blending site, and $n = 5$ tablets per batch. The sample means are given in Table 17.32.

TABLE 17.32 Sample means for Example 17.11

	Batch			
Site	1	2	3	Site Mean
1	5.082	4.84	5.134	5.01867
2	5.06	5.216	4.928	5.068
Overall mean				5.04333

From the data, we compute the following sums of squares:

$$\text{TSS} = (5.03 - 5.04333)^2 + (5.10 - 5.04333)^2 + \cdots + (5.07 - 5.04333)^2 = .76348$$

$$\text{SSA} = 15\{(5.01867 - 5.04333)^2 + (5.068 - 5.04333)^2\} = .01824$$

$$\text{SSB(A)} = 5\{(5.082 - 5.01867)^2 + (4.84 - 5.01867)^2 + (5.134 - 5.01867)^2 + (5.06 - 5.068)^2 + (5.216 - 5.068)^2 + (4.928 - 5.068)^2\} = .45401$$

$$\text{SSE} = \text{TSS} - \text{SSA} - \text{SSB(A)} = .76348 - .01824 - .45401 = .29123$$

The computer output for the analysis of this data set is given here. Note that the sums of squares differ slightly from our calculations. This is due to round-off error because we are dealing with very

small deviations. We will use the sums of squares from the computer output in the analysis of variance table for this experiment, which is given in Table 17.33.

TABLE 17.33
AOV table for experimental data

Source	SS	df	MS	*F*
A	.01825	1	.01825	.16
B(A)	.45401	4	.11350	9.39
Error	.29020	24	.01209	
Total	.76246	29		

```
CONTENT UNIFORMITY OF FILM-COATED TABLETS

General Linear Models Procedure

Dependent Variable: Y   CONTENT

Source                   DF   Sum of Squares          Mean     F Value     Pr > F
Model                     5       0.47226667    0.09445333        7.81     0.0002
Error                    24       0.29020000    0.01209167
Corrected Total          29       0.76246667

                  R-Square             C.V.      Root MSE              Y Mean
                  0.619393         2.180346       0.10996             5.04333

Source                   DF      Type III SS  Mean Square     F Value     Pr > F
SITE                      1       0.01825333   0.01825333        1.51     0.2311
BATH (SITE)               4       0.45401333   0.11350333        9.39     0.0001

Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: Y  CONTENT

Source: SITE
Error: MS(BATCH(SITE))
                           Denominator    Denominator
       DF    Type III MS            DF             MS     F Value     Pr > F
        1   0.0182533333             4   0.1135033333      0.1608     0.7089

Source: BATCH(SITE)
Error: MS(Error)
                           Denominator    Denominator
       DF     Type III MS            DF             MS     F Value     Pr > F
        4    0.1135033333            24   0.0120916667      9.3869     0.0001
```

b. and **c.** The F test for batches is

$$F = \frac{\text{MSB(A)}}{\text{MSE}} = 9.39$$

based on $df_1 = 4$ and $df_2 = 24$. Because the observed value of F, 9.39, exceeds the tabled value of F for $\alpha = .05, 2.78$, we conclude that there is considerable batch-to-batch variability in the content uniformity of the tablets. This test does not depend on whether the batches are random. ■

By now, you may have realized that a whole new series of experimental designs have opened up with the introduction of nested effects. Thinking beyond the two-factor design, one could imagine a general multifactor design with factor A, factor B nested in levels of factor A, factor C nested in levels of factors A and B, and so on. The analysis of variance table for a three-factor nested design with all factors random is shown in Table 17.34.

TABLE 17.34
AOV table for a three-factor nested design—all factors random (n observations per cell)

Source	SS	df	MS	EMS
A	SSA	$a - 1$	MSA	$\sigma_\varepsilon^2 + n\sigma_{\gamma(\tau,\beta)}^2 + cn\sigma_{\beta(\tau)}^2 + bcn\sigma_\tau^2$
B(A)	SSB(A)	$a(b - 1)$	MSB(A)	$\sigma_\varepsilon^2 + n\sigma_{\gamma(\tau,\beta)}^2 + cn\sigma_{\beta(\tau)}^2$
C(A, B)	SSC(A, B)	$ab(c - 1)$	MSC(A, B)	$\sigma_\varepsilon^2 + n\sigma_{\gamma(\tau,\beta)}^2$
Error	SSE	$abc(n - 1)$	MSE	σ_ε^2
Total	TSS	$abcn - 1$		

Other extensions of these designs are possible as well. For example, one could have a three-factor experiment in which factors A and B are cross-classified but factor C is nested within levels of factors A and B. This would be an example of a *partially nested design*.

Suppose that a marketing research firm is responsible for sampling potential customers to obtain their opinions on two products (A_1 and A_2) in four geographic areas of the country ($B_1, \ldots, B_4$). A random sample of six stores selling product A_i is obtained in each geographic area. For each store selected for product A_i in geographic area B_j, 10 people are interviewed concerning product A_i. For this design, factor C (stores) would be nested in levels of factors A (products) and B (geographic areas), and there would be $n = 10$ observations (opinions) for each level of factor C (stores) nested in levels of factors A and B. Factors A and B are fixed and crossed, while factor C is random.

17.7 RESEARCH STUDY: Factors Affecting Pressure Drops Across Expansion Joints

A major problem in power plants is that of pressure drops across expansion joints in electric turbines. The process engineer wants to design a study to identify the factors that are most likely to influence the pressure drop readings. Once these factors are identified and the most crucial factors are determined by the sizes of their contributions to the pressure drops across the expansion joints during the study, the engineer can make design changes in the process or alter the method by which the operators of the process are trained. These types of changes may be expensive or time consuming, so the engineer wants to be certain which factors will have the greatest impact on reducing the pressure drops.

Designing the Data Collection

The process engineer considered the following issues in designing an appropriate experiment to evaluate pressure drop:

1. What factors should be used in the study?
2. What levels of the factors are of interest?
3. How many levels are needed to adequately identify the important sources of variation?
4. How many replications per factor–level combination are needed to obtain a reliable estimate of the variance components?

5. What environmental factors may affect the performance of the pressure gauge during the test period?
6. What are the valid statistical procedures for evaluating the causes of the variability in pressure drops across the expansion joints?
7. What type of information should be included in a final report to document that all important sources of variability have been identified?

The factors selected for study were the gas temperature on the inlet side of the joint and the type of pressure gauge used by the operator. The engineer wants to know if the differences in gauge performance are affected by the temperature and hence decides that a factorial experiment is required to determine which of these factors has the greatest effect on the pressure drop. Three temperatures that cover the feasible range for operation of the turbine are 15°C, 25°C, and 35°C. There are hundreds of different types of pressure gauges used to monitor the pressure in the lines. Four types of gauges are randomly selected from the list of possible gauges for use in the study. In order to obtain a precise estimate of the mean pressure drop for each of the 12 factor–level combinations, it was decided to obtain six replications of each of 12 treatments. The data from the 72 experimental runs were given in Section 17.1.

A profile plot of the 12 sample treatment means is presented in Figure 17.2. From the plot, the mean pressure drops for gauge type G1 have larger changes over the observed temperature range than do the other three gauge types. In order to determine if this observed difference is more than just random variation, we will develop models and analysis techniques in the remainder of this chapter to enable us to identify which factors have the greatest contribution to the overall variation in the pressure drops.

The objective of the study was to determine if the pressure drops across the expansion joints in electric turbines were related to gas temperature. Also, the engineer wanted to assess the variation in readings from the various types of pressure gauges and determine whether variation in readings was consistent across different gas temperatures. In Table 17.1, we observed that there was a

FIGURE 17.2
Profile plot of mean pressure drop for the 12 gauge–temperature treatments

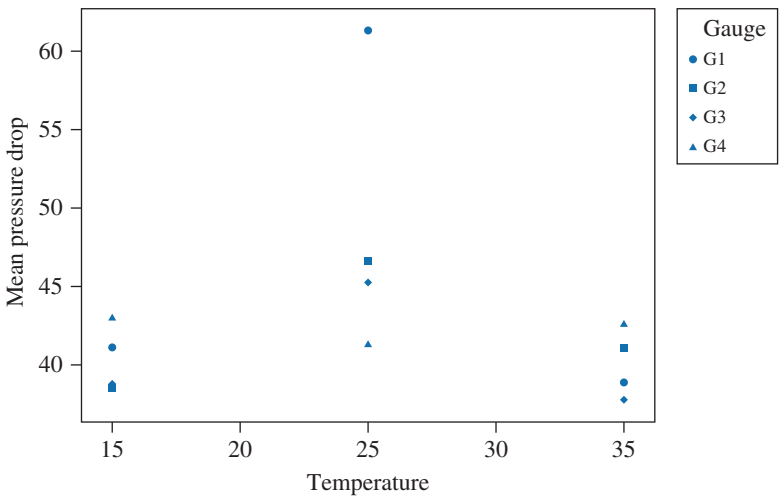


TABLE 17.35
Means and standard deviations of pressure drops readings

	Mean				Standard Deviation			
Temperature	**G1**	**G2**	**G3**	**G4**	**G1**	**G2**	**G3**	**G4**
15	41.17	38.50	38.67	43.00	3.31	3.62	3.72	4.69
25	61.33	46.67	45.33	41.33	4.27	3.44	2.07	4.97
35	39.00	41.17	37.83	42.67	4.69	2.79	3.31	4.18

slight increase in pressure drop as the temperature increased from 15°C to 25°C but a subsequent decrease in pressure drop when the temperature was further increased from 25°C to 35°C. The pressure drops recorded by the four gauges were fairly consistent over the three temperatures, with the exception that gauge G1 recorded a much higher mean pressure drop than the other three gauges at 25°C. Table 17.35, means and standard deviations for the 12 temperature–gauge combinations, reveals a fairly constant standard deviation, but gauge G1 had a much higher mean pressure drop at 25°C than the mean pressure drops of the other 11 temperature–gauge treatments.

Analyzing the Data

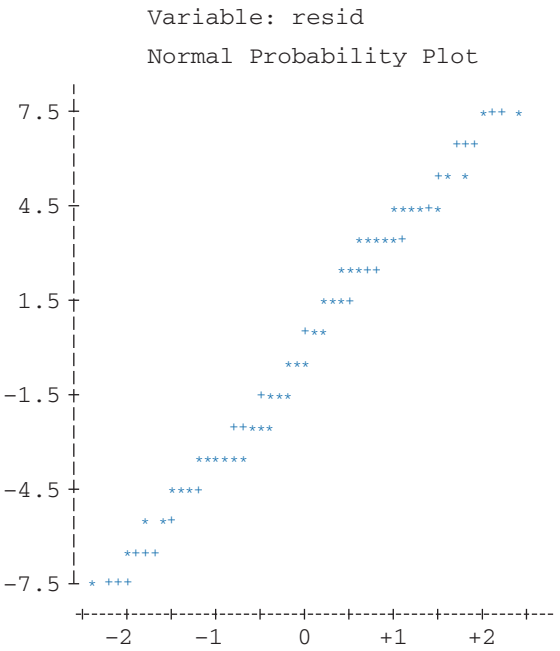
Since the four gauges were a random sample from a population of gauges available on the market, the gauge factor is a random effect. Thus, we want to assess whether the patterns observed in Table 17.35 and in Figure 17.2 were significant differences relative to the population from which the gauges were selected. Also, we want to determine if there are significant differences in mean pressure drops across the selected population. Additionally, we want to determine if there are significant differences in mean pressure drops across the temperature range 15°C to 35°C. The temperature factor has a fixed effect. The following model will be fit to the data:

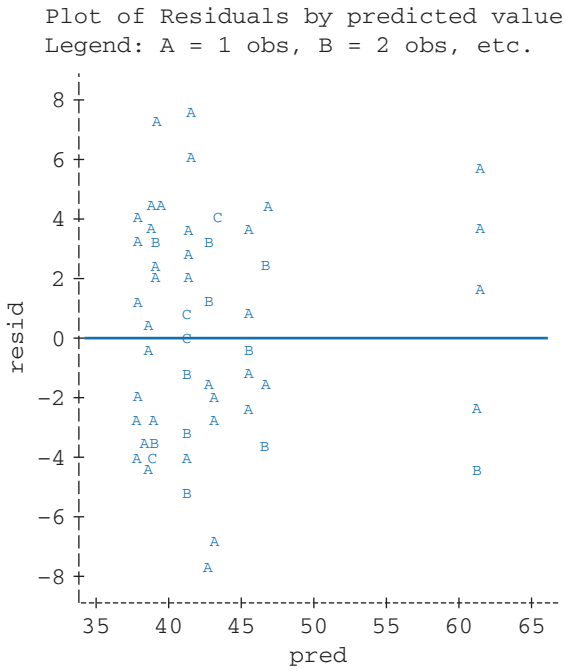
$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the pressure drop during the kth replication using gauge k with temperature i, τ_i is the fixed effect due to the ith temperature, β_j is the random effect due to the jth type of gauge, and $\tau\beta_{ij}$ is the interaction effect of the jth type of gauge observed under the ith temperature. Prior to running tests of hypotheses or constructing confidence intervals, we will evaluate the conditions that the experiment must satisfy in order for inferences to be appropriate. An examination of the following plots of the residuals will assist us in checking on the validity of the model conditions.

```
                    The UNIVARIATE Procedure
                        Variable: resid

                              Moments

N                          72    Sum Weights                72
Mean                        0    Sum Observations            0
Std Deviation      3.53254487    Variance           12.4788732
Skewness           0.01497112    Kurtosis           -0.8796331
```

```
              Tests for Normality

Test              --Statistic---    -----p Value------

Shapiro-Wilk      W    0.970843     Pr < W       0.0928

 Stem Leaf                Number           Boxplot
      7 07                   2                |
      6                                       |
      5 78                   2                |
      4 0002335              7                |
      3 023335777            9              +--|--+
      2 03338                5              |     |
      1 23378                5              |     |
      0 57788                5              |  +  |
     -0 533222               6              *-----*
     -1 877322               6              |     |
     -2 87330                5              |     |
     -3 877775330            9              |     |
     -4 5332000              7              +-----+
     -5 32                   2                 |
     -6                                        |
     -7 70                   2                 |
        ----+----+----+----+
```





The boxplot and stem-and-leaf plot of the residuals do not indicate any extreme values. The normal probability plot indicates a few residuals somewhat deviant from the fitted line. However, the test of normality yields a p-value of .0655, so there is not significant evidence that the residuals are not normally distributed. The plot of the residuals versus predicted values does not indicate a violation of the equal variances of the residuals assumption, since the spread in the residuals remains reasonably constant across the predicted values. Also, the table of standard deviations for the 12 treatments has values that are not very different in size. Thus, the conditions of normality and equal variance appear to be satisfied by the data. The condition that the gauges be randomly selected from a population of gauges and that the experimental runs be conducted in such a manner that the responses are independent would be checked through discussions with the process engineer concerning the manner in which the experiments were conducted. We will now present the AOV table, with notation T = temperature and G = type of gauge, as Table 17.36.

TABLE 17.36
AOV table for research study

Source	SS	df	MS	EMS	F	p-value
T	1,133.78	2	556.89	$\sigma_\varepsilon^2 + 6\sigma_{\tau\beta}^2 + 24\theta_\tau$	3.07	.1205
G	437.22	3	145.74	$\sigma_\varepsilon^2 + 6\sigma_{\tau\beta}^2 + 18\sigma_\beta^2$	.79	.5421
T*G	1,106.78	6	184.46	$\sigma_\varepsilon^2 + 6\sigma_{\tau\beta}^2$	12.49	$< .0001$
Error	886.00	60	14.77	σ_ε^2		
Total	3,563.78	71				

The computer output from fitting the model to the data is given here.

```
Dependent Variable: y DROP

                                      Sum of
       Source                DF         Squares   Mean Square   F Value   Pr > F
       Model                 11      2677.777778    243.434343     16.49   <.0001
       Error                 60       886.000000     14.766667
       Corrected Total       71      3563.777778

                  R-Square     Coeff Var      Root MSE      y Mean
                  0.751387      8.925078      3.842742    43.05556

       Source                DF     Type III SS   Mean Square   F Value   Pr > F
       t                      2      1133.777778    566.888889     38.39   <.0001
       g                      3       437.222222    145.740741      9.87   <.0001
       t*g                    6      1106.777778    184.462963     12.49   <.0001

       Source                DF     Type III SS   Mean Square   F Value   Pr > F
       t                      2      1133.777778    566.888889      3.07   0.1205
       g                      3       437.222222    145.740741      0.79   0.5421
       Error: MS(t*g)         6      1106.777778    184.462963

       Source                DF     Type III SS   Mean Square   F Value   Pr > F
       t*g                    6      1106.777778    184.462963     12.49   <.0001
       Error: MS(Error)      60       886.000000     14.766667
```

From the AOV table, we determine that there is a significant (p-value $< .0001$) interaction between the gas temperature and the type of gauge. Thus, the relationship between mean pressure drop and gas temperature across the temperature range 15°C to 35°C is not the same for all types of gauges. This conclusion is a confirmation of the relationship we observed in the profile plot given in Figure 17.2 for the four gauges used in the study. There is not a significant (p-value $= .5421$) difference in mean pressure drops due to the type of gauge. Thus, averaged over the temperature range used in the study, the gauges used to measure pressure drops are not significantly different with respect to mean pressure drop. Similarly, the mean pressure drops across the three temperatures were not significantly (p-value $= .1205$) different. Thus, the process engineer would conclude that the impact on pressure drop of the type of gauge varied depending on the temperature of the gases.

17.8 Summary

Fixed, random, and mixed models are easily distinguished if we think in terms of the general linear model. The fixed-effects model relates a response to $k \geq 1$ independent variables and one random component, whereas a random-effects model is a general linear model with $k = 0$ and more than one random component. The mixed model, a combination of the fixed- and the random-effects models, relates a response to $k \geq 1$ independent variables and more than one random component.

We illustrated the application of random-effects models to experimental situations for the completely randomized design and for the $a \times b$ factorial treatment structure laid off in a completely randomized design. We noted similarities between tests of significance in an analysis of variance for a random-effects model and for the corresponding fixed-effects model. Inferences resulting from an analysis of variance for a mixed model were illustrated using the $a \times b$ factorial treatment structure.

Unfortunately, in an introductory course, only a limited amount of time can be devoted to a discussion of random- and mixed-effects models. To expand

our discussion in the text, the results of Section 17.5 are useful in developing the expected mean squares for sources of variability in the analysis of variance table for balanced designs. Using these expectations, we can then attempt to construct appropriate test statistics for evaluating the significance of any of the fixed or random effects in the model.

The hardest part in any of these problems involving random- or mixed-effects models arises from trying to estimate $E(y)$, with an appropriate confidence interval for a random-effects model and the average value of y at some level or combination of levels for fixed effects in a mixed model. We illustrated how to obtain an estimate of $E(y)$ for a random-effects model and how to construct an approximate confidence interval. The problem becomes even more complicated for mixed models.

The final topics covered in this chapter were nested designs. A brief introduction showed several variations on the basic factorial experiments discussed in Chapters 14 and 15 and in earlier sections of this chapter. The designs presented are only a few of the more common designs possible when considering nested effects in a multifactor experimental setting. The interested reader should consult the references at the end of this book to pursue these topics in more detail; in particular, Kuehl (2000) is an excellent reference.

17.9 Exercises

17.2 A One-Factor Experiment with Random Treatment Effects

Engin. **17.1** The process engineer for a large paint manufacturer is concerned about the consistency of an ingredient in the paint that determines the ability of the paint to resist fading. The paint has a specification of 5% by weight of the ingredient. She designed the following study to assess the consistency. Ten batches of paint, each consisting of 500 1-liter containers of paint, are randomly selected from the previous week's production. From each of the 10 batches, five containers of paint are selected, and a determination of the percentage of the ingredient is made. The following table contains the percentages from the 50 determinations.

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5	Batch 6	Batch 7	Batch 8	Batch 9	Batch 10
4.18	5.60	7.59	4.25	2.18	5.11	5.68	4.61	8.72	4.67
2.29	4.74	7.46	5.39	5.88	7.61	7.55	7.14	6.93	7.85
1.40	1.86	5.79	4.81	3.07	3.46	2.30	4.61	5.25	2.21
8.69	6.29	5.09	7.75	5.25	6.57	2.15	5.23	8.97	9.57
1.01	2.25	5.47	6.10	3.50	6.35	8.92	3.56	4.34	4.85

a. Write a random-effects model for this study, identifying all the terms in the model.
b. Run an analysis of variance for the data collected in this study. Test for a significant batch effect using $\alpha = 0.05$.
c. Estimate the variance components associated with batches (σ_a^2) and containers within batches (σ_ε^2). What proportion of the total variation in percentage of the fade prevention ingredient is due to the batch-to-batch variation?

Engin. **17.2** Suppose the process engineer of Exercise 17.1 wanted to estimate the average percentage of the fade protection ingredient in a randomly selected container of the paint.

a. Use the data presented in Exercise 17.1 to form a point estimate of the average percentage of the fade protection ingredient in a randomly selected container of the paint.
b. Place a 95% confidence interval on the average percentage of the fade protection ingredient in a randomly selected container of the paint.

Ag. **17.3** A rancher is interested in determining if the average daily gain in weight of calves depends on the bull that sired the calf. Consider the following two situations:

Scenario A: The rancher has only five bulls. The five bulls are mated with randomly selected cows, and the average daily gains in weight by the calves produced by the matings are recorded.

Scenario B: The rancher has hundreds of bulls and randomly selects five bulls for inclusion in the study. The five bulls are mated with randomly selected cows, and the average daily gains in weight by the calves produced by the matings are recorded.

The data are given here.

Bull 1	Bull 2	Bull 3	Bull 4	Bull 5
1.20	1.16	.75	.96	.99
1.39	1.08	1.12	1.16	.85
1.36	1.22	1.02	1.05	1.10
1.39	.97	1.08	1.00	1.03
1.22	1.17	.83	1.12	.94
1.31	1.12	.98	1.15	.89

a. Write an appropriate linear statistical model for both scenario A and scenario B, identifying all the terms in the models.

b. State the null and alternative hypotheses for testing for a bull effect for each of the two scenarios.

Ag. **17.4** Refer to Exercise 17.3.

a. For scenario B, randomly selected bulls, run an analysis of variance and test for a significant bull effect.

b. Estimate the variance components associated with bulls (σ_a^2) and individual calves within bulls (σ_ε^2). What proportion of the total variation in average daily weight gain is due to the bull-to-bull variation?

c. Place a 95% confidence interval on the average daily weight gain for a calf sired by a randomly selected bull.

17.3 Extensions of Random-Effects Models

Med. **17.5** Periodontal disease may play a role in many diseases, some of which were unknown previously. For example, a recent study of the failure of joint replacement prostheses due to aseptic loosening demonstrated a link with bacterial DNA that was also found in dental plaque. Therefore, it is crucial that methods of determining bacterial DNA in plaque have a high degree of reliability. A study was conducted to examine the variability in the chemical analyses for specified bacterial DNA content in plaque. The two major sources of variation selected for investigation were the person conducting the analysis and the subjects supplying the plaque. The researchers randomly selected five analysts from a large pool of experienced analysts and 10 female subjects (ages 18–20). Plaque was scraped from the entire dentition of each subject and divided into five samples. Each of the analysts was given an unmarked sample from each of the subjects. The analysts then made a determination of the DNA content (in micrograms) for each of the 10 samples. The data are given here.

	Subject									
Analyst	1	2	3	4	5	6	7	8	9	10
1	9.9	10.6	11.5	11.3	10.5	8.0	10.6	12.2	8.0	9.7
2	10.2	10.6	11.3	11.6	10.3	8.2	10.7	12.8	7.9	9.6
3	10.1	10.5	11.1	11.3	10.1	7.9	10.4	12.6	7.7	9.3
4	10.2	10.5	11.2	11.3	10.2	7.9	10.5	12.7	7.8	9.4
5	10.4	10.9	11.4	11.6	10.6	8.4	10.9	12.5	8.1	9.5

a. Why do you think the researchers selected only female subjects who were essentially the same age?
b. Write an appropriate linear statistical model identifying all terms in the model.
c. State the null and alternative hypotheses for testing for an effect due to analyst.

Med. **17.6** Refer to Exercise 17.5.
a. Write down the expected mean squares.
b. Run an analysis of variance and test for a significant analyst effect.
c. Estimate the variance components associated with analysts, subjects, and error. What proportion of the total variation is associated with each of the three sources of variation?
d. Place a 95% confidence interval on the average amount of DNA in the plaque of a randomly selected subject.

Bus. **17.7** Beer is pasteurized by subjecting it to processes in manufacturing and packaging that attempt to kill, inactivate, or remove all yeast cells or other microorganisms, thereby preventing any further fermentation or microbiological decomposition of the packaged beer that might otherwise take place. Pasteurization impacts both the safety of the product and, more important, the taste of the beer. Therefore, in order to guarantee that the pasteurization has been effectively implemented, beer manufacturers have well-defined testing procedures. A large beer manufacturer has numerous breweries and is concerned about the variability in the effectiveness of the pasteurization process across its many facilities. Preliminary studies indicated that the manufacturer's many testing laboratories had varying ability to accurately determine the level of contamination in the beer. The manufacturer's quality control staff decided to concentrate its efforts on examining the variability in the level of contamination due to the effectiveness of the pasteurization processes and the variability due to the laboratory's determination of level of contamination.

The manufacturer's research staff designed the following study. Six laboratories are selected at random from the manufacturer's many breweries. Ten different pasteurization processes are randomly selected, and 12 samples of beer are selected from each of these processes. Two samples from each process are then sent to each laboratory. The laboratories count the microorganisms in each sample. The beer samples are coded so that the laboratories do not know which pasteurization process had treated the beer. The counts (units per μl) from the 10 laboratories are given here.

	Process									
Lab	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1,055	1,768	1,500	1,875	1,758	1,172	996	1,134	544	124
	1,056	1,763	1,474	1,883	1,762	1,215	994	1,120	590	176
2	2,390	2,202	958	2,664	2,614	2,029	1,516	1,982	113	1,555
	2,406	2,233	968	2,716	2,688	2,115	1,546	1,947	119	1,504
3	2,641	1,998	2,651	3,094	1,178	1,553	1,200	2,138	1,528	1,405
	2,721	2,067	2,718	3,124	1,159	1,517	1,190	2,179	1,531	1,384
4	1,508	1,090	1,380	1,394	1,777	1,399	1,709	1,848	1,064	904
	1,533	1,042	1,355	1,367	1,695	1,423	1,604	1,894	1,023	909
5	1,493	1,970	1,192	2,090	1,858	1,420	1,460	1,542	1,514	1,117
	1,448	1,999	1,164	2,096	1,891	1,415	1,439	1,527	1,587	1,067
6	2,633	1,098	1,466	2,063	1,884	1,896	932	1,888	1,247	595
	2,613	1,077	1,624	2,070	1,888	1,945	890	1,964	1,172	601

a. Write an appropriate linear statistical model, identifying all terms in the model.
b. Write down the expected mean squares.
c. State the null and alternative hypotheses for testing for an interaction effect, an effect due to laboratory, and an effect due to process.

Bus. **17.8** Refer to Exercise 17.7.

- **a.** Run an analysis of variance and test for significant effects.
- **b.** Estimate the variance components associated with interaction, laboratory, process, and error. What proportion of the total variation is associated with each of the four sources of variation?
- **c.** Is the effect due to laboratory or to process greater?

17.4 Mixed-Effects Models

Basic **17.9** A researchers wants to design a study to investigate two factors. He has the funding to investigate four levels of factor A and six levels of factor B, using 10 subjects randomly assigned to each of the 24 combinations of the two factors.

- **a.** Explain the difference in how the study would be conducted when both factors have random levels and when both factors have fixed levels.
- **b.** Explain the difference in the types of inferences that. can be made from a study with both factors having random levels and from a study with both factors having fixed levels.
- **c.** Explain the difference in the types of inferences that can be made from a study in which factor A has fixed levels and factor B has random levels and from a study in which both factors have fixed levels.

Env. **17.10** The following study was designed to evaluate the effectiveness of four chemicals developed to control fire ants. The type of environmental conditions in which the chemical is placed might affect the effectiveness of the treatment to kill fire ants. Thus, the researcher randomly selected five locations from a large selection of locations, with location representing a randomly selected environment. To reduce the effect of the different colonies of fire ants and the types of mounds they inhabit, the researcher created 40 artificial fire ant mounds and populated them with 50,000 ants having similar ancestry. The researcher randomly assigned 2 mounds to each of the 20 treatment–location combinations. The numbers of fire ants killed during a 1-week period were recorded. The numbers of fire ants killed (in thousands) are given here.

	Chemicals			
Locations	**1**	**2**	**3**	**4**
1	7.2	4.2	9.5	5.4
	9.6	3.5	9.3	3.9
2	8.5	2.9	8.8	6.3
	9.6	3.3	9.2	6.0
3	9.1	1.8	7.6	6.1
	8.6	2.4	7.1	5.6
4	8.2	3.6	7.3	5.0
	9.0	4.4	7.0	5.4
5	7.8	3.7	9.2	6.5
	8.0	3.9	8.3	6.9

- **a.** Write an appropriate linear statistical model for this study. Identify all the terms in your model.
- **b.** Compute the sum of squares for this experiment, and report this value in an AOV table. Be sure to include the expected mean squares column in the AOV table.

Env. **17.11** Refer to Exercise 17.10.

- **a.** Display a complete analysis of variance table including F tests and p-values.
- **b.** Is there a significant interaction between locations and chemicals? If the interaction is significant, what can we conclude about the effect of the chemicals?
- **c.** Are the main effects of chemicals and locations significant?

d. Based on your tests in parts (b) and (c), what can you say about the effect of chemicals on the number of fire ants killed?
e. What proportion of the variation in the number of fire ants killed can be attributed to chemicals, locations, interaction, and all other sources?
f. Are the conditions necessary to conduct the analysis in parts (b) and (c) satisfied? Justify your answer using the residuals.

Supplementary Exercises

Basic **17.12** A completely randomized design is conducted with five levels of factor A randomly selected from a population of levels and three levels of factor B the only levels of interest to the researcher. The experiment will be implemented by randomly assigning three experimental units to the 15 treatments obtained by crossing the levels of factors A and B.

a. Write a linear statistical model for this experiment. Identify all the terms in your model, and state all the conditions that are imposed on these terms.
b. Display a partial AOV table, including degrees of freedom and expected mean squares for all sources of variation.
c. For each of the main effects and interactions, display the ratio of mean squares that would be the appropriate F statistic for testing the significance of each of the terms.

Basic **17.13** A completely randomized design is conducted with five levels of factor A randomly selected from a population of levels, six levels of factor B randomly selected from a population of levels, and three levels of factor C the only levels of interest to the researcher. The experiment will be implemented by randomly assigning 10 experimental units to the 90 treatments obtained by crossing the levels of factors A, B, and C.

a. Write a linear statistical model for this experiment. Identify all the terms in your model, and state all the conditions that are imposed on these terms.
b. Display a partial AOV table, including degrees of freedom and expected mean squares for all sources of variation.
c. For each of the main effects and interactions, display the ratio of mean squares that would be the appropriate F statistic for testing the significance of each of the terms.

Basic **17.14** A completely randomized design is conducted with four levels of factor A randomly selected from a population of levels, three levels of factor B randomly selected from a population of levels, and five levels of factor C randomly selected from a population of levels. The experiment will be implemented by randomly assigning five experimental units to the 60 treatments obtained by crossing the levels of factors A, B, and C.

a. Write a linear statistical model for this experiment. Identify all the terms in your model, and state all the conditions that are imposed on these terms.
b. Display a partial AOV table, including degrees of freedom and expected mean squares for all sources of variation.
c. For each of the main effects and interactions, display the ratio of mean squares that would be the appropriate F statistic for testing the significance of each of the terms.

Basic **17.15** A completely randomized design is conducted with three levels of factor A randomly selected from a population of levels; six levels of factor B, which are the only levels of interest to the researcher; and three levels of factor C, which are the only levels of interest to the researcher. The experiment will be implemented by randomly assigning three experimental units to the 54 treatments obtained by crossing the levels of Factors A, B, and C.

a. Write a linear statistical model for this experiment. Identify all the terms in your model, and state all the conditions that are imposed on these terms.
b. Display a partial AOV table, including degrees of freedom and expected mean squares for all sources of variation.
c. For each of the main effects and interactions, display the ratio of mean squares that would be the appropriate F statistic for testing the significance of each of the terms.

Env. **17.16** Refer to Exercise 17.10. Suppose the four chemicals were randomly selected from the hundreds of different chemicals used to control fire ants. The researcher was interested in whether the effectiveness of a chemical to control fire ants varied across different environments.

a. Write an appropriate model for this situation. Indicate how the conditions placed on the terms in the model differ from the conditions placed on the model used when the chemicals were the only chemicals of interest to the researchers.

b. Construct the AOV table and test all relevant hypotheses.

c. Compare the conclusions and inferences in this problem to those of Exercise 17.10.

Env. **17.17** Refer to Exercise 17.16.

a. Which model and analysis seem to be more appropriate? Explain your answer.

b. Under what circumstances would a fixed-effects model be appropriate?

Engin. **17.18** A university in a large urban area is having a major problem with traffic congestion. A study was funded to determine the volume of traffic on campus streets by cars that are not associated with university business. One small phase of the study involved obtaining daily counts on the number of cars crossing but not making use of campus facilities. Video cameras were placed at each entrance to the university. The license plate number and the time of entrance or exit for each car passing through the campus entrances were recorded. Using these data and allowing a reasonable time for cars to traverse the campus, the researchers were able to determine the number of cars crossing the campus but not making use of any campus facilities during the business day. A random sample of 12 weeks throughout the academic year was used in the analysis. During each of the 12 selected weeks, the traffic volume was recorded during the business hours of the five days. The data are given in the following table.

Week of Traffic Volume Count											
Wk1	**Wk2**	**Wk3**	**Wk4**	**Wk5**	**Wk6**	**Wk7**	**Wk8**	**Wk9**	**Wk10**	**Wk11**	**Wk12**
680	438	539	264	693	530	700	518	427	368	579	210
656	487	601	198	646	575	636	497	534	305	580	250
597	496	578	195	652	548	610	510	501	347	536	219
643	518	609	258	638	561	652	452	485	367	567	268
656	491	558	231	682	546	687	461	492	353	592	197

a. Write an appropriate linear statistical model for this experiment. Identify all the terms in your model, and state all the conditions that are imposed on these terms.

b. Display a complete analysis of variance table, including expected mean squares, F tests, and p-values.

c. Estimate the variation in the mean weekly traffic volumes across the year using the information in the AOV table.

d. Is the variation in traffic volume within weeks greater or less than the variation in mean weekly variations across the year? Justify your answer using the estimators of the variance components obtained from the information in the AOV table.

e. Use the residuals from the fitted model to determine if there are any violations in the conditions necessary to conduct the tests of hypotheses in this experiment.

Gov. **17.19** The public safety department at a large urban university was concerned about criminal activities involving nonstudents stealing bicycles and laptops from students. The campus police designed a study to investigate the number of automobiles entering the campus that do not have a campus parking sticker or do not enter a campus parking facility. The police were suspicious that such individuals may be involved in criminal activities. A team of criminal justice students was stationed at each entrance to the campus to monitor simultaneously the license numbers of all cars and to determine if each car had a campus parking sticker. By utilizing the computer records of all campus parking facilities, which record the license number of all cars upon their entrance to a parking facility, the teams were able to determine the numbers of cars entering the campus but not using campus facilities. Data were collected during a random sample of 12 weeks throughout the academic year. The counts of "suspicious" cars were recorded on the five business days during the selected 12 weeks and appear here.

	Week											
Day	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Mon	52	51	52	54	56	54	51	56	51	48	52	53
Tue	47	50	50	51	55	51	49	54	49	46	51	50
Wed	49	50	50	52	54	51	49	54	49	47	52	50
Thu	49	50	49	52	54	50	48	54	49	46	51	51
Fri	44	48	48	50	53	50	48	52	48	45	50	51

a. Write an appropriate linear statistical model for this experiment. Identify all the terms in your model, and state all the conditions that are imposed on these terms.
b. Display a complete analysis of variance table, including expected mean squares, F tests, and p-values.
c. Bicycle and laptop thefts seem to occur in clusters. Therefore, if the count of "suspicious" cars is associated with theft, then there should be a large variation in the weekly counts. Does the number of suspicious cars arriving on campus on a weekly basis remain fairly constant over the academic year?
d. Use the residuals from the fitted model to determine if there are any violations in the conditions necessary to conduct the tests of hypotheses in this experiment.

Gov. **17.20** Refer to Exercise 17.19. Estimate the average number of "suspicious" cars entering the campus for a randomly selected week during the academic year, and include an appropriate confidence interval.

Med. **17.21** A study was designed to evaluate the effectiveness of new treatments to reduce the systolic blood pressure of patients determined to have high blood pressure. Three drugs were selected for evaluation (D1, D2, and D3). There are numerous nondrug treatments for reducing blood pressure, including various combinations of a controlled diet, exercise programs, biofeedback, and so on. The researchers randomly selected three nondrug treatments (ND1, ND2, and ND3) for examination in the study. The age of the patient often may hinder the effectiveness of any treatment. Thus, patients with high blood pressure were divided into two age groups (A1 and A2). A group of 54 patients was divided into the two age groups and then randomly assigned to a combination of one of the three drugs and one of the three nondrug treatments. After patients had participated in the program for 2 months, the reductions in their systolic blood pressure readings from their blood pressure readings at the beginning of the program were recorded. These values are given in the following table.

	Age A1			**Age A2**		
	Nondrug			**Nondrug**		
	ND1	**ND2**	**ND3**	**ND1**	**ND2**	**ND3**
Drug	33	37	41	34	48	44
D1	34	38	42	33	46	46
	35	36	39	38	45	49
Drug	46	44	43	47	44	44
D2	45	48	44	49	48	46
	46	49	45	45	46	41
Drug	38	45	36	36	46	38
D3	34	45	37	39	47	36
	37	44	35	35	44	35

a. Write a model for this study. Identify all the terms in your model, and state all the necessary conditions placed on the terms in the model.
b. Construct the AOV table for the study, including the expected mean squares.
c. Test the significance of all relevant sources of variation. Use $\alpha = .05$.

d. What conclusions do you draw about the differences in the effectiveness of the combinations of nondrug and drug treatments for high blood pressure?
e. Would it be appropriate to recommend a treatment based on these data? Justify your answer.

Engin. **17.22** Refer to Exercise 15.22. Suppose that we consider the five investigators to be a random sample from a population of all possible investigators for the rocket propellant experiment.
a. Write an appropriate linear statistical model, identifying all the terms and listing your assumptions.
b. Perform an analysis of variance. Include an expected mean squares column in the analysis of variance table.

Engin. **17.23** Refer to Exercise 17.22.
a. If the five investigators are considered to be a fixed effect, what are the hypotheses being tested, and what conclusions can be drawn if the null hypothesis is rejected? State your answer in terms of a parameter(s) reflecting the difference in the propellants.
b. If the five investigators are considered to be a random effect, what are the hypotheses being tested, and what conclusions can be drawn if the null hypothesis is rejected? State your answer in terms of a parameter(s) reflecting the difference in the propellants.

17.24 Refer to Exercise 14.33. Suppose that the two laboratories were randomly selected from a population of laboratories for participation in the study, which also included time as a possible source of variability.
a. Obtain the expected mean squares for all sources of variability.
b. Test all relevant sources of variability for significance. Use $\alpha = .05$.
c. Compare the results obtained here to the results obtained in Exercise 14.34.
d. Does considering the laboratory effects to be random effects seem more relevant than considering them to be fixed effects? Explain your answer.

17.25 Refer to Exercise 14.31. Suppose that the five pane designs were randomly selected from a population of pane designs for participation in the study.
a. Obtain the expected mean squares for all sources of variability.
b. Test all relevant sources of variability for significance. Use $\alpha = .05$.
c. Compare the results obtained here to the results obtained in Exercise 14.31.
d. Does considering the pane design effects to be random effects seem more relevant than considering them to be fixed effects? Explain your answer.

17.26 Refer to the study described in Exercise 14.27.
a. Considering the nine medications to be randomly selected from a population of possible medications, write a model for the study.
b. Give the expected mean squares for all sources of variability.
c. Indicate how your analysis and conclusions would change from those of Exercise 14.27.

Engin. **17.27** The two most crucial factors that influence the strength of solders used in cementing computer chips into the mother board of the guidance system of an airplane are identified as the machine used to insert the solder and the operator of the machine. Four solder machines and three operators were randomly selected from the many machines and operators available at the company's plants. Each operator made two solders on each of the four machines. The resulting strength determinations of the solders are given here.

	Machine			
Operator	**1**	**2**	**3**	**4**
1	204	205	203	205
	205	210	204	203
2	205	205	206	209
	207	206	204	207
3	211	207	209	215
	209	210	214	212

a. Write a model for this study. Include all the terms and conditions placed on the terms in the model.
b. Present the AOV table for this study, and include the expected mean squares.
c. What conclusions can you draw about the effect of machine and operator on the variability in solder strength?

17.28 Refer to Exercise 17.27.
a. Estimate the variance components in this study.
b. Proportionally allocate the sources of variability with respect to the total variability in solder strength.
c. Place a 95% confidence interval on the average solder strength.

Env. **17.29** Core soil samples are taken in each of six locations within a territory being investigated for surface mining of bituminous coal. Each of the core samples is divided into four subsamples for separate analyses of the sulfur content of the sample.
a. Identify the design and give a model for this experimental setting.
b. Give the sources of variability and degrees of freedom for an AOV.

17.30 The sample data for Exercise 17.29 are shown here. Run an AOV and draw conclusions. Use $\alpha = .05$.

	Analysis			
Location	**1**	**2**	**3**	**4**
1	15.2	16.8	17.5	16.2
2	13.1	13.8	12.6	12.9
3	17.5	17.1	16.7	16.5
4	18.3	18.4	18.6	17.9
5	12.8	13.6	14.2	14.0
6	13.5	13.9	13.6	14.1

Engin. **17.31** Tablet hardness is one comparative measure for different formulations of the same drug product; some combinations of ingredients (in addition to the active drug) in a formulation give rise to harder tablets than do other combinations. Suppose that three batches of a formulation are randomly selected for examination. Three different 1-kg samples of tablets are randomly selected from each batch, and seven tablets are randomly selected for testing from each of the 1-kg samples. The hardness readings are given here.

	Batch 1			Batch 2			Batch 3		
Sample	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
	85	76	95	108	117	101	71	81	72
	94	87	98	100	106	108	85	70	68
	91	90	94	105	103	100	78	84	80
	98	91	96	109	109	99	68	83	72
	85	88	99	104	100	117	85	72	75
	96	94	100	102	104	109	67	81	79
	93	96	93	108	102	105	76	78	74

a. Identify the design.
b. Give an appropriate model with assumptions.
c. Give the sources of variability and degrees of freedom for an AOV.
d. Perform an analysis of variance, and draw conclusions about the tablet hardness data for the formulation under study. Use $\alpha = .05$.

Sci. **17.32** An anthropologist is interested in the impact of the usage of mind-altering drugs in religious ceremonies. She selects five underdeveloped countries for inclusion in her study. She then selects 10 tribes in each country. Finally, she randomly selects 20 families from each tribe for an in-depth interview. After the interview, the anthropologist assigns a score that reflects the impact of the usage of mind-altering drugs in religious ceremonies. The researcher is interested in determining if there is a difference in the average scores across countries and what the degree of variability is in the index across tribes and families. In this study, there are three factors of interest to the researcher: country, tribe, and family.

- **a.** Identify each of the factors as fixed or random; justify your answer.
- **b.** State whether the factors are nested or crossed; provide reasons for your answers.
- **c.** Provide an AOV table that includes source of variation, df, and expected mean squares.

Sci. **17.33** A soil scientist is studying the potassium content of three major soil types in Texas. For each of the three soil types, the scientist randomly selects five sites in which this soil type is the dominant soil type within the site. Within each site, five soil samples are randomly selected, and the potassium content is determined. The soil scientist is interested in the level of difference in the average potassium contents across the three soil types and in the degree of variability in potassium contents within sites.

- **a.** Identify each of the factors as fixed or random; justify your answer.
- **b.** State whether the factors are nested or crossed; provide reasons for your answers.
- **c.** Provide an AOV table that includes source of variation, df, and expected mean squares.

Edu. **17.34** There has been a major initiative to include the use of laptop computers as a part of the lesson plan in math and science courses in middle schools. There has been some resistance to the inclusion due to costs and the reluctance on the part of some teachers to increase their use of technology-based instruction. A major study was designed in a large midwestern state to study these issues. The school districts in the state were divided into three groups: urban, rural, and mixed urban-rural. Ten school districts were randomly selected within each of these three groups. Five randomly selected schools provided a weeklong workshop on how to include laptops in their daily instruction, and the other five schools were given only a manual that described laptop implementation strategies. Six teachers were randomly selected from each of the 30 schools. The teachers' classroom and lesson plans were then examined to determine the degree to which they had included laptops into their instruction. The researchers were interested in determining the impact on instruction of type of school district and type of training. Also, they wanted to measure the variability among schools of the same type and among teachers from the same schools.

- **a.** Identify each of the factors as fixed or random; justify your answer.
- **b.** State whether the factors are nested or crossed; provide reasons for your answers.
- **c.** Provide an AOV table that includes source of variation, df, and expected mean squares.

Med. **17.35** ***The following study is from Oehlert (2000).*** Dental fillings made from gold can vary in hardness depending on how the metal is treated prior to its placement in the tooth. Two factors thought to influence the hardness are the gold alloy and the condensation method. In addition, some dentists performing the dental work are better at some types of filling than others. Five dentists were randomly selected and agreed to participate in the experiment. Each dentist prepared 24 fillings (in random order), one for each of the combinations of condensation method (three levels) and type of alloy (eight levels). The levels of condensation and type of alloy are the only levels of interest to the researchers. The fillings were then measured for hardness using the Diamond Pyramid Hardness Number (big scores are better). The data are contained in the following table:

		Alloy							
Dentist	**Method**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	1	792	824	813	792	792	907	792	835
1	2	772	772	782	698	665	1115	835	870
1	3	782	803	752	620	835	847	560	585

		Alloy							
Dentist	**Method**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
2	1	803	803	715	803	813	858	907	882
2	2	752	772	772	782	743	933	792	824
2	3	715	707	835	715	673	698	734	681
3	1	715	724	743	627	752	858	762	724
3	2	792	715	813	743	613	824	847	782
3	3	762	606	743	681	743	715	824	681
4	1	673	946	792	743	762	894	792	649
4	2	657	743	690	882	772	813	870	858
4	3	690	245	493	707	289	715	813	312
5	1	634	715	707	698	715	772	1048	870
5	2	649	724	803	665	752	824	933	835
5	3	724	627	421	483	405	536	405	312

a. Write an appropriate linear statistical model for this experiment. Identify all the terms in your model, and state all the conditions that are imposed on these terms.

b. Display a complete analysis of variance table, including expected mean squares, F tests, and p-values.

c. Is there significant evidence of an interaction between condensation method and type of alloy?

Med. **17.36** Refer to Exercise 17.35.

a. Group the types of alloys such that alloys within a group have similar mean hardness scores.

b. Group the types of condensation methods such that alloys within a group have similar mean hardness scores.

e. Estimate the variation in the mean hardnesses due to the dentist.

e. Use the residuals from the fitted model to determine if there are any violations in the conditions necessary to conduct the tests of hypotheses in this experiment.

Health **17.37** A state health department conducted an experiment to evaluate the reliability of assessing the level of contamination of *e. coli* in three food sources, meat, fruit, and vegetables. There are four unique methods for assessing *e. coli*—M1, M2, M3, and M4—and hundreds of laboratories that use one or more of theses methods in the United States. For each of the methods of assessment, five laboratories are randomly selected to participate in the study. Forty containers are prepared for each food source by spiking the container with a known level of contamination of *e. coli* and then placing the container in a controlled climate for 3 weeks to allow the *e. coli* level to stabilize. Six containers, two of each of the three food sources, are then sent to each of the 20 laboratories selected for the study. The *e. coli* level (cfu/g), Y_{ijkl}, determined by the kth lab using assessment method j for the lth container of food source i is recorded for each of the 120 containers. The health department wants to compare the mean *e. coli* levels of the four assessment methods and their differences across the food sources. It also wants to determine if there are major differences in the mean *e. coli* determinations across the many laboratories in the United States.

	Assessment Method																			
	M1					**M2**					**M3**					**M4**				
	Lab					Lab					Lab					Lab				
Source	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	L16	L17	L18	L19	L20
Meat	12.3	13.2	12.9	13.2	12.9	14.5	14.3	14.5	14.3	14.5	14.4	13.5	14.7	13.5	14.7	14.8	16.4	15.6	14.8	15.2
	12.6	13.0	13.0	14.0	13.9	15.0	15.6	14.8	15.6	14.8	14.1	13.4	14.6	13.4	14.2	15.3	14.3	16.4	15.4	14.4
Fruit	13.2	14.4	12.9	14.1	12.8	13.2	14.2	14.4	13.3	14.5	13.4	14.5	12.7	13.5	12.7	12.2	14.4	12.4	13.4	13.2
	13.4	14.5	13.7	14.1	13.4	14.2	13.4	14.6	13.6	13.8	14.1	14.4	14.2	14.4	14.2	13.3	13.6	13.8	13.8	13.3
Veg.	13.1	13.4	13.6	13.8	12.8	13.5	13.3	13.5	14.3	13.5	14.3	13.5	13.7	14.5	13.7	12.2	14.3	13.6	13.4	14.4
	12.5	14.0	13.0	14.1	13.3	14.0	12.6	12.8	13.6	12.8	15.1	15.4	15.2	15.4	15.2	13.3	13.9	12.7	13.9	14.1

a. Write a model that displays an appropriate relationship between a level of contamination, Y_{ijkl}, and its possible sources of variation. Include any restrictions on the parameters in your model and any distributional properties of the random variables in your model.
b. Do the necessary conditions for testing hypotheses and constructing confidence intervals appear to be satisfied? Justify your answers.
c. Construct an ANOVA table for this experiment. Make sure to include expected mean squares and the p-values for the F tests.
d. At the $\alpha = .05$ level, which main effects and interaction effects are significant? Justify your answer by including the relevant p-values.
e. What are your overall conclusions about the differences in the four assessment methods?

Health **17.38** Refer to Exercise 17.37.

a. For meat products, separate the four assessment methods into groups such that all assessment methods in a group are not significantly different from one another with respect to their mean *e. coli* levels. Use an experimentwise error rate of $\alpha = .05$.
b. For fruit products, separate the four assessment methods into groups such that all assessment methods in a group are not significantly different from one another with respect to their mean *e. coli* levels. Use an experimentwise error rate of $\alpha = .05$.
c. For vegetable products, separate the four assessment methods into groups such that all assessment methods in a group are not significantly different from one another with respect to their mean *e. coli* levels. Use an experimentwise error rate of $\alpha = .05$.
d. Provide a 95% confidence interval on the mean *e. coli* level of a container of meat for each of the assessment methods.
e. Provide a 95% confidence interval on the mean *e. coli* level of a container of fruit for each of the assessment methods.
f. Provide a 95% confidence interval on the mean *e. coli* level of a container of vegetables for each of the assessment methods.
g. Was it necessary to do a separate grouping of the assessment methods for each of the food types? Justify your answer based on the tests conducted in the AOV table.

CHAPTER 18

Split-Plot, Repeated Measures, and Crossover Designs

18.1 Introduction and Abstract of Research Study

In all of the experimental situations discussed so far in this text (except for the paired difference experiment), we have assumed that only one observation is taken on each experimental unit. For example, in an experiment to compare the effects of three different cardiovascular compounds on blood pressure, we could use a completely randomized design where n_1 patients are assigned to compound 1, n_2 to compound 2, and n_3 to compound 3. Then the model would be

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where τ_i is the fixed effect due to compound i and ε_{ij} is the random effect associated with patient j treated with compound i. For this design, we would get one measurement (y_{ij}) for each patient.

The practicalities of many applied research settings make it mandatory from a cost and efficiency standpoint to obtain more than one observation per experimental unit. For example, in conducting clinical research, it is often difficult to find patients who have the condition to be studied *and* who are willing to participate in a clinical trial. Hence, it is important to obtain as much information as possible once a suitable number of patients have been located.

TABLE 18.1
Repeated time points for each patient

Compound	Time Period 1	2	...	t
1	y_{111}	y_{112}	$\cdots$	y_{11t}
	$\vdots$	$\vdots$		$\vdots$
	y_{1n_11}	y_{1n_12}	$\cdots$	y_{1n_1t}
2	y_{211}	y_{212}	$\cdots$	y_{21t}
	$\vdots$	$\vdots$		$\vdots$
	y_{2n_21}	y_{2n_22}	$\cdots$	y_{2n_2t}
3	y_{311}	y_{312}	$\cdots$	y_{31t}
	$\vdots$	$\vdots$		$\vdots$
	y_{3n_31}	y_{3n_32}	$\cdots$	y_{3n_3t}

split-plot design

When the experiment involves a factorial treatment structure, the implementation of one of two factors may be more time consuming or more expensive or may require more material than the other factors. In circumstances such as these, a **split-plot design** is often implemented. For example, in an educational research study involving two factors, teaching methodologies and individual tutorial techniques, the teaching methodologies would be applied to the entire classroom of students. The tutorial techniques would then be applied to the individual students within the classroom. In an agricultural experiment involving the factors levels of irrigation and varieties of cotton, the irrigation systems must apply the water to large sections of land, which would then be subdivided into smaller plots. The different varieties of cotton would then be planted on the smaller plots. In both of these examples, the levels of one factor are applied to a large experimental unit, which is then subdivided into smaller units to which the levels of the second factor are then assigned.

crossover designed experiment

In a **crossover designed experiment,** each subject receives all treatments. The individual subjects in the study are serving as blocks and hence decreasing the experimental error. This provides an increased precision in the treatment comparisons when compared to the design in which each subject receives a single treatment. In the **repeated measures designed experiment,** we obtain t different measurements corresponding to t different time points following administration of the assigned treatment. This experimental setting is shown in Table 18.1. In Table 18.1, y_{ijk} denotes the observation at the time k for the jth patient on compound i. Note that we are getting $t > 1$ observations per patient, rather than only 1.

repeated measures designed experiment

The multiple observations over time on the same subject often yield a more efficient use of experimental resources than using a different subject for each observation time. Fewer subjects are required, with a subsequent reduction in cost. Also, the estimation of time trends will be measured with a greater degree of precision. The methods of this chapter can be used to analyze data from split-plot experiments, crossover studies, and repeated measures studies. The application of these designs is broad-based. Applications abound in the pharmaceutical industry and in the research and development (R & D) and manufacturing operations of most industries. Medical research, ecological studies, and numerous other areas of research involve the evaluation of time trends and hence may find the repeated measures design useful. An extension of these designs may also be appropriate for studies in which the data have a spatial relationship in place of the time trend. Examples include the reclamation of strip-mined coal fields, evaluation of the effects of an oil spill, and air pollution around an industrial facility. Studies involving spatially repeated measures are generally more complex to model than

the time trends we will address in this chapter. Further reading on the modeling of spatial data can be found in Ripley (1998), Haining (1990), and Cressie (1993).

The following research study will illustrate the evaluation of time trends in a repeated measures design.

Abstract of Research Study: Effects of Oil Spill on Plant Growth

We examined a small portion of this research study in Chapter 6. On January 7, 1992, an underground oil pipeline ruptured and caused the contamination of a marsh along the Chiltipin Creek in San Patricio County, Texas. The cleanup process consisted of burning the contaminated regions in the marsh. To evaluate the influence of the oil spill on the flora, the researchers designed a study of plant growth after the burn was finished. In an unpublished Texas A&M University dissertation, ***Newman (1998) describes the researchers' findings with respect to* Distichlis spicata,** a flora of particular importance to the area of the spill.

Two questions of importance to the researchers were as follows:

1. Did the oil site recover after the spill and burning?
2. How long did it take for the recovery?

To answer these questions, the researchers needed to have a baseline to which they could compare the *Distichlis spicata* density in the months after the burning of the site. The density of the flora depended on soil characteristics, slope of the land, environmental conditions, weather, and many other factors. The researchers designated as the control site a nearby section of land that was not affected by the oil spill but that had soil and environmental properties similar to those of the spill site. At both the oil spill site and the control site, 20 tracts were randomly chosen. After a 9-month transition period, measurements were taken at approximately 3-month intervals for a total of eight time periods. During each time period, the number of *Distichlis spicata* within each of the 40 tracts was recorded.

The experimental design is a repeated measures design with two treatments, the oil spill and the control region, and eight measurements taken over time on each of the tracts over a 2-year period. To answer the researchers' questions, we will state them in terms of the *Distichlis spicata* counts. Thus, our research hypotheses are stated as follows:

1. Was there a difference in the average density of *Distichlis spicata* between the oil spill tracts and the control tracts during the study period?
2. Were there significant trends in average density of *Distichlis spicata* during the study period?
3. Were the trends for the oil spill and control tracts different?

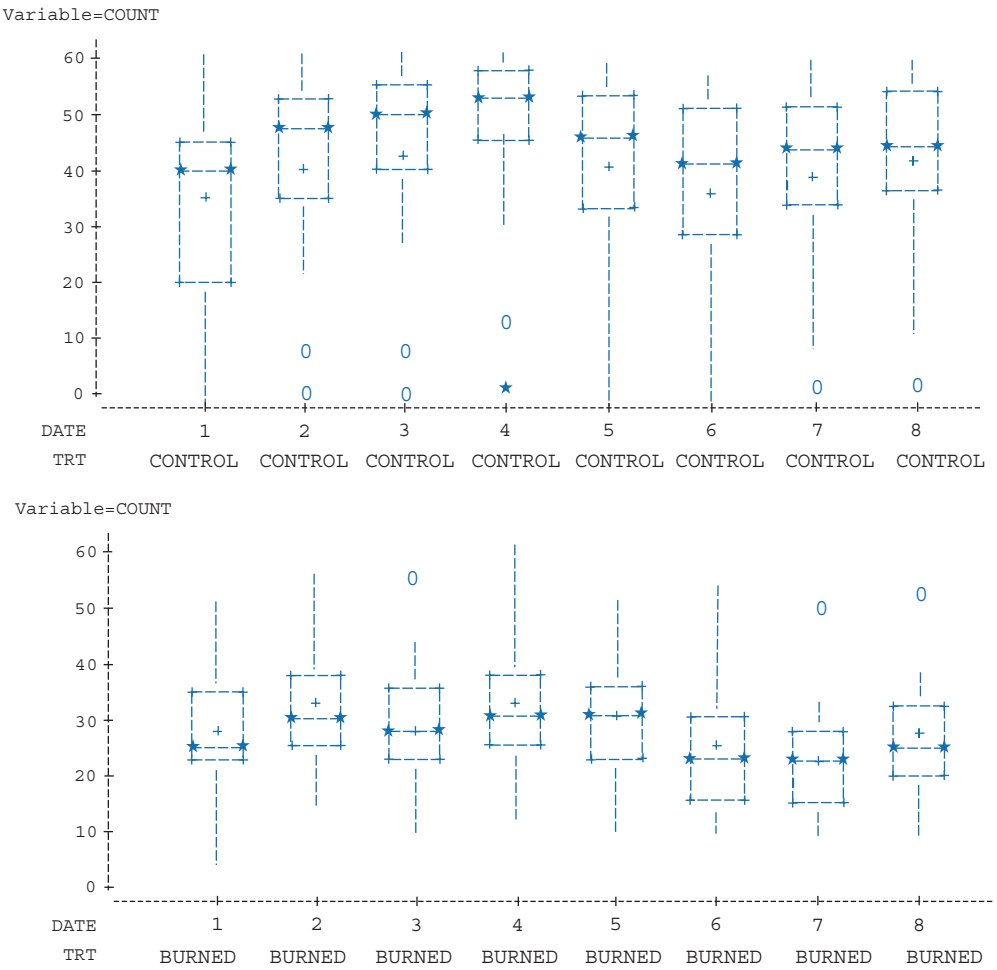
The data consisted of the number of *Distichlis spicata* plants found on each tract during the eight observation periods on both the control and the burned (oil spill) sites. There were a total of 320 data values. The data are given in Table 18.2.

The flora counts are plotted in Figure 18.1, using boxplots for each date and treatment. The boxplots reveal that the control plots have higher median flora counts than the oil spill plots. The control plots, however, are somewhat more variable than the oil spill plots. This may be due to the burning treatment used on the oil spill plots, which often results in more homogeneous tract that was conditions than were present prior to the burning. The extension of these observations beyond the tracts in the study to the population of tracts will require modeling of

TABLE 18.2
Number of *Distichlis spicata* under two treatments

Treatment	Tract	Oct. 92	Jul. 93	Oct. 93	Jan. 94	Apr. 94	Jul. 94	Oct. 94	Jan. 95
Burned	1	27	25	18	21	26	22	20	27
	2	5	15	10	12	10	11	12	9
	3	17	26	26	25	15	10	14	17
	4	41	41	42	38	34	26	26	25
	5	25	28	22	27	24	16	18	23
	6	11	24	13	20	16	13	10	14
	7	37	40	33	31	32	30	25	31
	8	38	38	33	38	39	35	32	38
	9	31	33	25	30	28	21	17	19
	10	24	25	21	24	24	19	17	22
	11	22	27	31	30	32	30	25	34
	12	26	45	39	35	35	36	30	27
	13	32	38	34	45	41	28	31	31
	14	35	37	35	42	35	32	27	29
	15	26	23	19	18	21	13	11	19
	16	22	29	24	24	20	16	18	24
	17	50	54	56	60	51	52	49	52
	18	17	29	23	39	31	24	26	34
	19	25	37	29	32	28	14	13	24
	20	33	39	39	48	36	34	30	34
Control	1	7	0	0	1	0	0	0	0
	2	57	46	49	51	48	43	40	40
	3	43	59	59	60	58	53	55	58
	4	43	53	52	53	53	53	52	54
	5	59	55	59	60	54	47	54	53
	6	42	48	50	48	43	37	38	38
	7	35	42	50	55	41	40	44	45
	8	40	51	53	57	53	38	43	36
	9	24	52	54	59	57	55	57	39
	10	42	49	50	54	51	44	39	41
	11	16	31	39	47	24	22	33	35
	12	54	58	60	60	54	51	48	51
	13	30	43	43	47	39	36	49	56
	14	47	50	60	60	54	52	57	57
	15	40	40	47	49	43	41	48	52
	16	11	23	27	31	17	19	24	29
	17	41	45	42	44	41	33	31	42
	18	50	52	55	53	45	42	35	51
	19	8	8	7	12	6	5	8	10
	20	0	0	0	1	0	0	0	0

FIGURE 18.1
Boxplots of flora counts by treatment and date



the data and testing of the relevant statistical research hypotheses. We will provide this analysis at the end of the chapter after introducing the methods of analyzing repeated measures designs.

18.2 Split-Plot Designed Experiments

Split-plot designs are another type of experimental design that can be used to implement studies involving factorial treatment structures. The split-plot design is generally implemented when one or more of the factors is more time consuming, expensive, or difficult to apply to the experimental units than the other factors. The major difference between split-plot designs and completely randomized designs is that split-plot designs have more than one randomization when assigning treatments to experimental units and the experimental units for the levels of one factor are different from the experimental units for the other factors. Split-plot designs originated in agricultural experimentation. We will illustrate the split-plot design with an example involving soybeans.

The yields of three different varieties of soybeans are to be compared under two different levels of fertilizer application. If we were interested in getting (say) $n = 2$ observations at each combination of fertilizer and variety of soybeans, we

would need 12 equal-sized plots. Taking fertilizers as factor A and varieties as a treatment factor T, one possible design would be an 2×3 factorial treatment structure in a completely randomized design with $n = 2$ observations per factor–level combination. However, since the application of fertilizer to a plot occurs when the soil is being prepared for planting, it would be difficult (logistically) to first apply fertilizer A_1 to six of the plots dictated by the factorial arrangement of factors A and T and then fertilizer A_2 to the other six plots before planting the required varieties of soybeans in each plot.

An easier design to execute would have each fertilizer applied to two larger "wholeplots" and then the varieties of soybeans planted in three "subplots" (equal in size to the plots of the previous design) within each wholeplot. A design of this type appears in Figure 18.2.

This design is called a split-plot design, and with this design, there is a two-stage randomization. First, levels of factor A (fertilizers) are randomly assigned to the wholeplots; second, the levels of factor T (soybeans) are randomly assigned to the subplots within a wholeplot (see Figure 18.3). Using this design, it would be much easier to prepare the soil and to apply the appropriate fertilizer to the larger wholeplots and then to plant varieties of soybeans in the subplots rather than preparing the soil and applying fertilizer to the subplots and then planting soybeans in the subplots, as would be the case for a standard 2×3 factorial experiment.

Because the randomization at the wholeplot level and at the subplot level is according to a completely randomized design, the design is often referred to as a completely randomized split-plot design.

FIGURE 18.2
Split-plot design

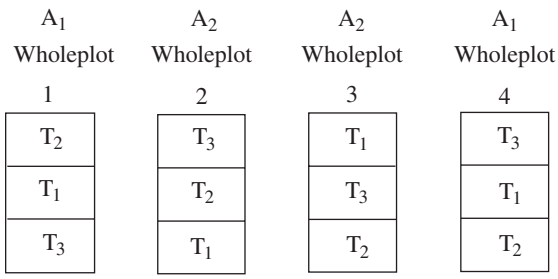


FIGURE 18.3
Two-stage randomization for a completely randomized split-plot design

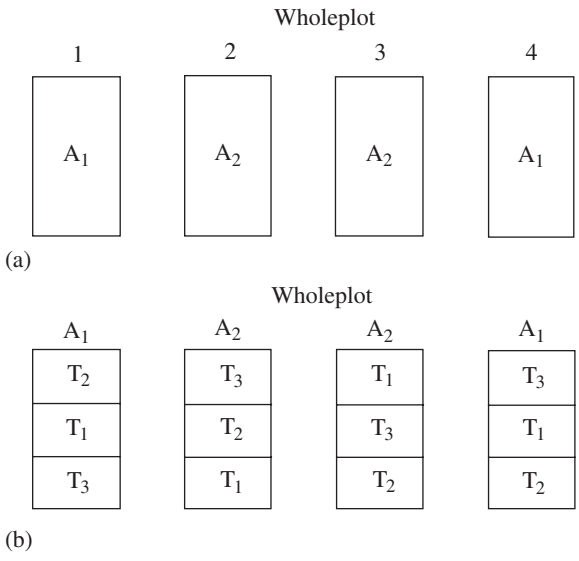


TABLE 18.3
AOV for a completely randomized split-plot design

Source	SS	df	EMS
A	SSA	$a - 1$	$\sigma_\varepsilon^2 + t\sigma_\delta^2 + tn\theta_\tau$
Wholeplot error	SS(A)	$a(n - 1)$	$\sigma_\varepsilon^2 + t\sigma_\delta^2$
T	SST	$t - 1$	$\sigma_\varepsilon^2 + an\theta_\gamma$
AT	SSAT	$(a - 1)(t - 1)$	$\sigma_\varepsilon^2 + n\theta_{\tau\gamma}$
Subplot error	SSE	$a(n - 1)(t - 1)$	σ_ε^2
Total	TSS	$atn - 1$	

Consider the model for the completely randomized split-plot design with a levels of factor A, t levels of factor T, and n repetitions of the ith level of factor A. If y_{ijk} denotes the kth response for the ith level of factor A and the jth level of factor T, then

$$y_{ijk} = \mu + \tau_i + \delta_{k(i)} + \gamma_j + \tau\gamma_{ij} + \varepsilon_{ijk}$$

where

τ_i: Fixed effect for ith level of A.

γ_j: Fixed effect for jth level of T.

$\tau\gamma_{ij}$: Fixed effect for ith level of A and jth level of T.

$\delta_{k(i)}$: Random effect for the kth wholeplot receiving the ith level of A. The δ_{ik} are independent and normal with mean 0 and variance σ_δ^2.

ε_{ijk}: Random error. The ε_{ijk} are independent and normal with mean 0 and variance σ_ε^2.

The $\delta_{k(i)}$ and ε_{ijk} are mutually independent. The AOV for this model and design is shown in Table 18.3.

You could compute the sums of square for the AOV using our standard formulas, but we suggest going to computer output to get them. It follows from the expected mean square that we have the following analyses:

Wholeplot Analysis

$$H_0\text{: } \theta_\tau = 0 \text{ (or, equivalently, } H_0\text{: All } \tau_i = 0),\ F = \frac{\text{MSA}}{\text{MS(A)}}$$

Subplot Analysis

$$H_0\text{: } \theta_{\tau\gamma} = 0 \text{ (or, equivalently, } H_0\text{: All } \tau\gamma_{ij} = 0),\ F = \frac{\text{MSAT}}{\text{MSE}}$$

$$H_0\text{: } \theta_\gamma = 0 \text{ (or, equivalently, } H_0\text{: All } \gamma_j = 0),\ F = \frac{\text{MST}}{\text{MSE}}$$

A variation on this design introduces a *blocking factor* (such as farms). Thus, for our example, there may be $b = 2$ farms with $a = 2$ wholeplots per farm and $t = 3$ subplots per wholeplot. This design is shown in Figure 18.4. Because the randomization to the wholeplots is done according to a randomized block design and the randomization to the subplot units within a wholeplot occurs according to a completely randomized design, the design is often referred to as a randomized block split-plot design.

The model for this more general two-factor split-plot design laid off in b blocks is as follows:

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \gamma_k + \tau\gamma_{ik} + \varepsilon_{ijk}$$

FIGURE 18.4
Randomized block split-plot design

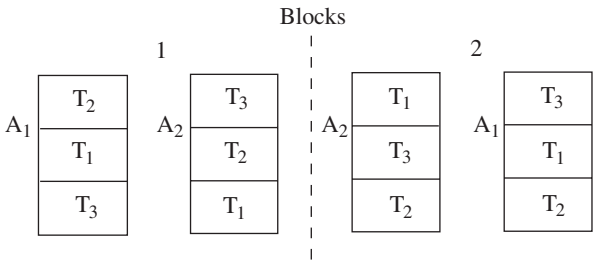


TABLE 18.4
AOV for a randomized block split-plot design (A, T fixed; blocks random)

Source	SS	df	EMS
Blocks	SSB	$b-1$	$\sigma_\varepsilon^2 + at\sigma_\beta^2$
A	SSA	$a-1$	$\sigma_\varepsilon^2 + t\sigma_{\tau\beta}^2 + bt\theta_\tau$
AB (wholeplot error)	SSAB	$(a-1)(b-1)$	$\sigma_\varepsilon^2 + t\sigma_{\tau\beta}^2$
T	SST	$(t-1)$	$\sigma_\varepsilon^2 + ab\theta_\gamma$
AT	SSAT	$(a-1)(t-1)$	$\sigma_\varepsilon^2 + b\theta_{\tau\gamma}$
Subplot error	SSE	$a(b-1)(t-1)$	σ_ε^2
Totals	TSS	$abt-1$	

where y_{ijk} denotes the measurement receiving the ith level of factor A and the kth level of factor T in the jth block. The parameters τ_i, γ_k, and $\tau\gamma_{ik}$ are the usual main effects and interaction parameters for a two-factor experiment, whereas β_j is the effect due to block j and $\tau\beta_{ij}$ is the interaction between the ith level of factor A and the jth block. The analysis corresponding to this model is shown in Table 18.4. Here we assume factors A and T are fixed effects, whereas blocks are random.

The sums of squares for the sources of variability listed in Table 18.4 can be obtained using the general formulas for main effects and interactions in a factorial experiment or from an appropriate software package. Using these expected mean squares, we can obtain a valid F test for factor A in the wholeplot portion of the analysis and for factor T and the AT interaction in the subplot portion. These are shown here. Note that no test is made for the variability due to blocks.

Wholeplot Analysis

$$H_0\text{: } \theta_\tau = 0 \text{ (or, equivalently, } H_0\text{: All } \tau_i = 0), F = \frac{\text{MSA}}{\text{MSAB}}$$

Subplot Analysis

$$H_0\text{: } \theta_{\tau\gamma} = 0 \text{ (or, equivalently, } H_0\text{: All } \tau\gamma_{ik} = 0), F = \frac{\text{MSAT}}{\text{MSE}}$$

$$H_0\text{: } \theta_\gamma = 0 \text{ (or, equivalently, } H_0\text{: All } \gamma_k = 0), F = \frac{\text{MST}}{\text{MSE}}$$

EXAMPLE 18.1

Soybeans are an important crop throughout the world. They are planted for use as both an oil and a source for protein. The vast majority of the crop is used for vegetable oil or defatted soy meal, which is then used for feed for various farm animals. To a much lesser extent, soybeans are consumed directly as food by humans. However, soybean products are an ingredient in a wide variety of processed foods. A study was designed to determine if additional phosphorus applied to the soil

would increase the yield of soybean. There are three major varieties of soybeans of interest (V_1, V_2, and V_3) and four levels of phosphorus (0, 30, 60, and 90 pounds per acre). The researchers have nine plots of land available for the study, which are grouped into blocks of three plots each based on the soil characteristics of the plots. Because of the complexities of planting the soybeans on plots of the given size, it was decided to plant a single variety of soybeans on each plot and then divide each plot into four subplots. The researchers randomly assigned a variety to one plot within each block of three plots and then randomly assigned the levels of phosphorus to the four subplots within each plot. The yields (bushels/acre) from the 36 plots are given in Table 18.5.

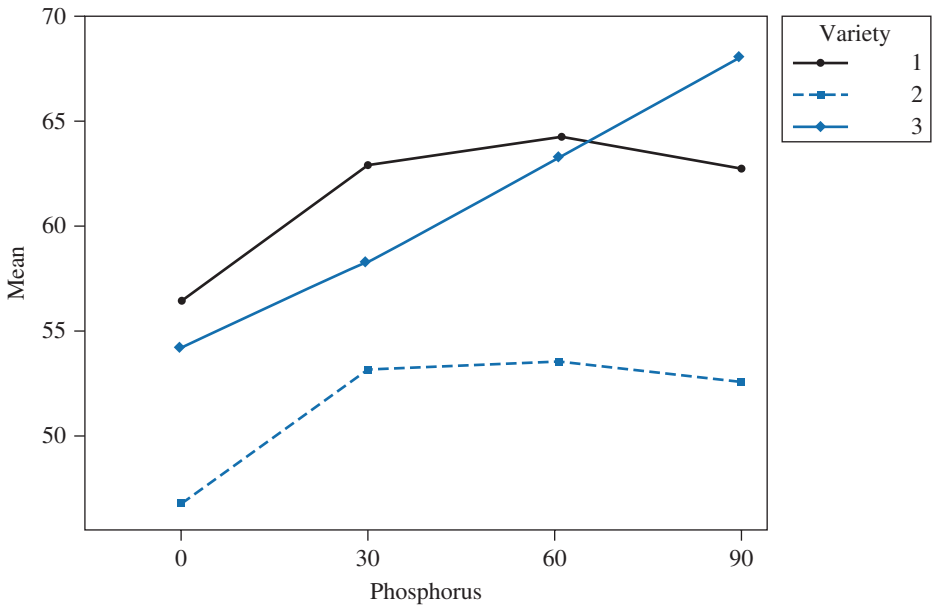
TABLE 18.5
Soybean yield data

	Block								
	B1			B2			B3		
Phosphorus	V_1	V_2	V_3	V_1	V_2	V_3	V_1	V_2	V_3
0	53.5	44.8	50.7	62.2	52.5	61.4	53.4	43.1	50.6
30	60.6	51.0	54.9	68.8	58.7	64.9	59.5	49.6	54.8
60	60.8	51.5	59.4	70.9	59.4	70.0	61.0	49.7	60.5
90	59.6	49.9	64.7	67.8	58.1	74.4	60.3	49.5	65.0

Conduct an analysis of variance using the sample data. Test whether there is an increase in the average yield with increasing amounts of phosphorus and whether the relationship between average yield and amount of phosphorus applied to the fields is the same for the three varieties.

Solution For this study, we have a randomized complete block design with a split-plot structure. Variety, with three levels, is the wholeplot treatment, and amount of phosphorus is the split-plot treatment. A profile plot of the interaction between variety and phosphorus level is given in Figure 18.5.

FIGURE 18.5
Interaction plot for soybean experiment



From the plot it would appear that the relationship between average yield and amount of phosphorus for variety V_3 is different from the relationship for the other two varieties.

The output from SAS is given here.

```
  Class Level Information

Class      Levels   Values
B               3   1 2 3
V               3   1 2 3
P               4   0 30 60 90

Number of Observations Read     36

Dependent Variable: y

                                  Sum of
Source                DF         Squares   Mean Square   F Value   Pr > F
Model                 17     1967.406389    115.729788    510.99   <.0001
Error                 18        4.076667      0.226481
Corrected Total       35     1971.483056

Source                DF     Type III SS   Mean Square   F Value   Pr > F
V                      2     763.2505556   381.6252778   1685.02   <.0001
B                      2     671.8072222   335.9036111   1483.14   <.0001
B*V                    4       6.5627778     1.6406944      7.24   0.0012
P                      3     408.3719444   136.1239815    601.04   <.0001
V*P                    6     117.4138889    19.5689815     86.40   <.0001

Source             TYPE III Expected Mean Square
V                  Var(Error) + 4 Var(B*V) + Q(V,V*P)
B                  Var(Error) + 4 Var(B*V) + 12 Var(B)
B*V                Var(Error) + 4 Var(B*V)
P                  Var(Error) + Q(P,V*P)
V*P                Var(Error) + Q(V*P)

Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: y

   Source             DF     Type III SS   Mean Square   F Value   Pr > F
*  V                   2      763.250556    381.625278    232.60   <.0001
   B                   2      671.807222    335.903611    204.73   <.0001

   Error: MS(B*V)      4        6.562778      1.640694
*  This test assumes one or more other fixed effects are zero.

   Source             DF     Type III SS   Mean Square   F Value   Pr > F
   B*V                 4        6.562778      1.640694      7.24   0.0012
*  P                   3      408.371944    136.123981    601.04   <.0001
   V*P                 6      117.413889     19.568981     86.40   <.0001

   Error: MS(Error)   18        4.076667      0.226481
*  This test assumes one or more other fixed effects are zero.
```

We can summarize the information from the SAS output into the following analysis of variance table, Table 18.6, with the following notation: B = blocks, V = variety, and P = phosphorus.

TABLE 18.6
AOV table for soybean experiment

Source	df	SS	MS	*F*	*p*-value
B	2	763.25	381.63	204.73	< .0001
V	2	671.81	335.90	232.60	< .0001
BV(wholeplot error)	4	6.56	1.64	7.24	.0012
P	3	408.37	136.12	601.04	< .0001
PV	6	117.41	19.57	86.40	< .0001
Subplot error	18	4.08	0.23		
Total	35	1,971.48			

It is important to note in the SAS output that the first set of values in the AOV table used MSE as the divisor for all F tests. Further down in the SAS output, the correct tests are conducted. The results from the AOV table confirm our observations from the profile plot. There is a significant variety by phosphorus interaction from which we can conclude that the relationships between average yield and amount of phosphorus added to the soil are not the same for the three varieties. In fact, for varieties V_1 and V_2, the average yield increases as the amount of phosphorus increases up to a phosphorus level of 60 but appears to remain at this level for a subsequent increase in phosphorus. The relationship for variety V_3 shows that the average yield continues to increase when the level of phosphorus is increased from 60 to 90. The next step in the analysis would be to conduct a multiple comparison of the variety means at each level of phosphorus or to examine the significance of various trends in the average yields for increasing phosphorus levels separately for each variety. ■

The distinction between this two-factor split-plot design and the standard two-factor experiments discussed in Chapter 14 lies in the randomization. In a split-plot design, there are two stages to the randomization process; first, levels of factor A are randomized to the wholeplots within each block, and then levels of factor B are randomized to the subplot units within each wholeplot of every block. In contrast, for a two-factor experiment laid off in a randomized block design (see Section 15.4), the randomization is a one-step procedure; treatments (factor–level combinations of the two factors) are randomized to the experimental units in each block. The post-AOV analysis involving mean separations, contrasts, estimated treatment means, and confidence intervals are somewhat more complex for the split-plot design than for the designs that we have discussed previously. Excellent references for further reading on this topic are Kuehl (2000), Snedecor and Cochran (1980), and Oehlert (2000).

18.3 Single-Factor Experiments with Repeated Measures

In Section 18.1, we discussed some reasons why one might want to get more than one observation per patient. Another reason for obtaining more than one observation per patient is that frequently the variability *among* or *between* patients is much greater than the variability *within* a patient. We observed this in the paired t test example of Section 6.4. If this is the case, it might be better to block on patients and to give each patient each treatment. Then the comparison among compounds is a within-patient comparison rather than a comparison between patients, as would be the case with the single-factor experiment with n_i different patients assigned to compound i. A single-factor design that reflects this within-patient emphasis is shown in Table 18.7.

TABLE 18.7
A within-patient comparison of compounds 1, 2, and 3

	Patient			
Compound	**1**	**2**	**...**	***n***
1	y_{11}	y_{12}	...	y_{1n}
2	y_{21}	y_{22}	...	y_{2n}
3	y_{31}	y_{32}	...	y_{3n}

With this design, the three compounds are administered in sequence to each of the n patients. A compound is administered to a patient during a given treatment period. After a sufficiently long "washout" period, another compound is given to the same patient. This procedure is repeated until the patient has been treated with all three compounds. The order in which the compounds are administered is randomized. In this design, it is crucial that the washout period between treatments be sufficiently long that the results for one compound do not affect the results for another compound.

Another effect that may need to be considered is the time period in which the response was recorded. A period effect is not a change in the response due to the treatment but a change in the response that would have occurred even in the absence of the treatment. Period effects, when they occur, are often a reflection of a variety of influences. For example, the period effect may be associated with seasonal effects, changes in conditions under which the measurements are obtained, a progression of the disease, or psychological effects of the application of multiple treatments. The experiment described in Table 18.7 would not permit the estimation of a period effect because the various treatment sequences are randomly assigned to the patients. If there was the possibility of period effects being present, then the investigator would randomly assign patients to the sequences (six possible sequences in Table 18.7) in such a way that there would be an equal number of patients for each of the sequences. We will discuss this type of design in Section 18.5.

Here again we are obtaining more than one observation per patient and presumably getting more useful information about the three drug products in question. One model for this experimental setting is

$$y_{ij} = \mu + \tau_i + \delta_j + \varepsilon_{ij}$$

where μ is the overall mean response, τ_i is the effect of the ith compound, δ_j is the effect of the jth patient, and ε_{ij} is the experimental error for the jth patient receiving the ith compound.

Note that this model looks like any other single-factor experimental setting with a compounds and n patients. However, the assumptions are different because we are obtaining more than one observation per patient. For this model, we make the following assumptions.

1. The τ_is are constants with $\tau_a = 0$.
2. The δ_j are independent and normally distributed $(0, \sigma_\delta^2)$.
3. The ε_{ij}s are independent of the δ_js.
4. The ε_{ij}s are normally distributed $(0, \sigma_\varepsilon^2)$.
5. The ε_{ij}s have the following correlation relationships:
   ε_{ij} and $\varepsilon_{i'j}$ are correlated for $i \neq i'$.
   ε_{ij} and $\varepsilon_{i'j'}$ are independent for $j \neq j'$.

That is, two observations from the same patient are correlated, but observations from different patients are independent. From these assumptions, it can be shown that the variance of y_{ij} is $\sigma_\delta^2 + \sigma_\varepsilon^2$. A further assumption is that the covariance for any two observations from patient j, y_{ij} and $y_{i'j}$, is constant. These assumptions give rise to a variance–covariance matrix for the observations, which exhibits *compound symmetry*. The discussion of correlated observations is beyond the scope of this book, and we refer the interested reader to Kuehl (2000) and Vonesh and Chinchilli (1997).

The analysis of variance for the experimental design being discussed and this set of assumptions is shown in Table 18.8. This AOV should be familiar. When the

TABLE 18.8
AOV for the experimental setting depicted in Table 18.7

Source	SS	df	EMS (A fixed, patients random)
Patients	SSP	$n - 1$	$\sigma_\varepsilon^2 + a\sigma_\delta^2$
A	SSA	$a - 1$	$\sigma_\varepsilon^2 + n\theta_\tau$
Error	SSE	$(a - 1)(n - 1)$	σ_ε^2
Totals	TSS	$an - 1$	

assumptions hold, and hence when compound symmetry holds, the statistical test on factor A (F = MSA/MSE) is appropriate. However, there are some other, more general conditions that also lead to a valid F test for factor A using F = MSA/MSE. How restrictive are these assumptions, and how can we tell when the test is appropriate?

There are no easy answers to these questions because there are no simple tests to check for compound symmetry. The general conditions (called the Huynh–Feldt conditions) under which the F test for factor A is valid are often not met because observations on the same patient taken closely in time are more highly correlated than are observations taken further apart in time. So be careful about this. In general, when the variance–covariance matrix does not follow a pattern of compound symmetry, the F test for factor A has a positive bias, which allows rejection of H_0: All $\tau_i = 0$ more often than is indicated by the critical F-values.

From a practical standpoint, the best thing to do in a given experimental setting is to make certain that there is sufficient time between applications of the treatment to allow washout (or elimination) of the previous treatment and to make certain that the design is applied in only those situations where the disease is relatively stable, so that following treatment and washout each patient (or experimental unit) is essentially the same as prior to receiving treatment. For example, even when studying the effect of blood-pressure-lowering drugs, we would expect the hypertension to be stable enough that the patients would return to their predrug blood pressure levels after washout of the first assigned compound before receiving the second assigned compound, and so on.

In Section 18.4, more will be said about how to judge whether the underlying assumptions for the test hold and, if they do not, how to proceed. For further information on this topic, refer to higher-level textbooks covering repeated measures experiments in detail (for example, Kuehl, 2000, and Vonesh and Chinchilli, 1997).

EXAMPLE 18.2

An exercise physiologist designed a study to evaluate the impact of the steepness of running courses on the peak heart rate (PHR) of well-conditioned runners. There are four 5-mile courses that have been rated as flat, slightly steep, moderately steep, and very steep with respect to the general steepness of the terrain. The 20 runners will run each of the four courses in a randomly assigned order. There will be sufficient time between the runs that there should not be any carryover effect, and the weather conditions during the runs will be essentially the same. Therefore, the researcher felt confident that the model $y_{ij} = \mu + \tau_i + \delta_j + \varepsilon_{ij}$ would be an appropriate model for analyzing the difference in the mean peak heart rates over the four courses. The mean heart rates are given in Table 18.9.

TABLE 18.9
Mean heart rate data

	Slope					Slope			
Runner	Flat	Slight	Moderate	Steep	Runner	Flat	Slight	Moderate	Steep
1	133	143	155	154	11	132	145	146	157
2	138	136	142	154	12	132	134	144	146
3	133	149	154	151	13	128	127	137	138
4	128	144	143	150	14	119	132	138	139
5	130	139	136	145	15	127	132	140	138
6	139	152	152	163	16	129	134	140	154
7	123	129	131	142	17	137	138	149	155
8	128	132	142	148	18	123	132	145	140
9	109	137	122	128	19	120	137	139	142
10	143	151	161	160	20	129	143	140	139

Determine if there is a significant difference in the mean heart rates of runners over the four degrees of steepness. Estimate the variation in the heart rates associated with runner and model error. The following output was obtained from SAS.

```
The GLM Procedure

                        Class Level Information

Class         Levels    Values

S                  4    Flat Moderate Slight Steep

R                 20    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Dependent Variable: y

                                    Sum of
Source                     DF       Squares   Mean Square   F Value   Pr > F
Model                      22   7667.675000    348.530682     18.34   <.0001
Error                      57   1083.512500     19.008991
Corrected Total            79   8751.187500

R-Square     Coeff Var      Root MSE      y Mean
0.876187      3.129604      4.359930    139.3125

Source                     DF   Type III SS   Mean Square   F Value   Pr > F

S                           3   3619.237500   1206.412500     63.47   <.0001
R                          19   4048.437500    213.075658     11.21   <.0001

                    Differences of Least Squares Means

                                                 Standard
Effect    S         _S          Estimate           Error      DF     Adj P

S         Flat      Moderate   -13.8000        1.3787        57     <.0001
S         Flat      Slight      -9.3000        1.3781        57     <.0001
S         Flat      Steep      -18.1500        1.3787        57     <.0001
S         Moderate  Slight       4.5000        1.3787        57     0.0098
S         Moderate  Steep       -4.3500        1.3787        57     0.0133
S         Slight    Steep       -8.8500        1.3787        57     <.0001
```

Solution From the output, we have the p-value associated with the F test of

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{versus} \quad H_a: \text{Not all } \mu_i\text{s are equal.}$$

as p-value $< .0001$. Thus, we can conclude that there is significant evidence of a difference in the mean heart rates over the four levels of steepness. The Tukey–Kramer pairwise test for difference demonstrates that there is significant evidence of a difference in all pairs of means.

	Slope			
	Flat	**Slight**	**Moderate**	**Steep**
Mean	129.0	138.3	142.8	147.15
Grouping	*a*	*b*	*c*	*d*

The estimated variance components are given by

$$\hat{\sigma}^2_{\text{Error}} = \text{MSE} = 19.01$$

$$\hat{\sigma}^2_{\text{Runner}} = \frac{\text{MS}_{\text{Runner}} - \text{MSE}}{4} = 48.52$$

Therefore, 72% of the variation in the heart rates was due to the differences in runners and 28% was due to all other sources. ■

18.4 Two-Factor Experiments with Repeated Measures on One of the Factors

We can extend our discussion of repeated measures experiments to two-factor settings. For example, in comparing the blood-pressure-lowering effects of cardiovascular compounds, we could randomize the patients so that n different patients receive each of the three compounds. Repeated measurements occur due to taking multiple measurements across time for each patient. For example, we might be interested in obtaining blood pressure readings immediately prior to receiving a single dose of the assigned compound and then every 15 minutes for the first hour and hourly thereafter for the next 6 hours.

This type of experiment can be described as follows. There are m treatments with n experimental units randomly assigned to each of the treatments. Each experimental unit is assigned to a single treatment with t measurements taken on each of the experimental units. The data for this type of experiment are depicted in Table 18.10. Note that this is a two-factor experiment (treatments and time) with repeated measurements taken over the time factor.

The analysis of a repeated measures design can, under certain conditions, be approximated by the methods used in a split-plot experiment. Each treatment

TABLE 18.10
Measurements at t time points for each experimental unit

		Time Period			
Treatment	**Exper. Unit**	**1**	**2**	**...**	***t***
1	1	y_{111}	y_{112}	...	y_{11t}
	⋮	⋮	⋮	⋮	
	n	y_{1n1}	y_{1n2}	...	y_{1nt}
2	1	y_{211}	y_{212}	...	y_{21t}
	⋮	⋮	⋮	⋮	
	n	y_{2n1}	y_{2n2}	...	y_{2nt}
⋮	⋮	⋮	⋮	...	⋮
m	1	y_{m11}	y_{m12}	...	y_{m1t}
	⋮	⋮	⋮	⋮	
	n	y_{mn1}	y_{mn2}	...	y_{mnt}

is randomly assigned to an experimental unit, EU. This is the wholeplot in the split-plot design. Each EU is then measured at t time points. This is considered the split-plot unit. The major difference between split plots and repeated measures is that in a split-plot design the levels of factor A are randomly assigned to the wholeplot EUs and the levels of factor B are randomly assigned to the split-plot EUs, whereas in the repeated measures design the second randomization does not occur. The treatment (factor A) is randomly assigned to the EUs (wholeplot EUs), but the levels of factor B (time) are *not* randomly assigned to a subunit of the EU. Thus, there may be a strong correlation between the measurements across time (or location) for those measurements produced by the same EU.

Therefore, the split-plot analysis is an appropriate analysis for a repeated measures experiment only when the covariance matrix of the measurements satisfies a particular type of structure: **compound symmetry:**

compound symmetry

$$Cov(y_{ijk}, y_{i'j'k}) = \begin{cases} \sigma_\varepsilon^2 & \text{when } i = i', j = j' \\ \rho\sigma_\varepsilon^2 & \text{when } i = i', j \neq j' \\ 0 & \text{when } i \neq i' \end{cases}$$

where y_{ijk} is the measurement from the kth EU receiving treatment i at time j. Thus

$$Corr(y_{ijk}, y_{ij'k}) = Cor(y_{ijk}, y_{ij'k})/\sigma_\varepsilon^2 = \rho$$

This implies that there is a constant correlation between observations no matter how far apart they are taken in time. This may not be realistic in many applications. One would think that observations in adjacent time periods would be more highly correlated than observations taken two or three time periods apart.

However, if the compound symmetry condition is satisfied, then the split-plot analysis produces a relatively accurate approximation to the p-values for testing hypotheses about treatment, time, and interaction effects. In fact, a somewhat less restrictive condition is all that is required. The Huynh–Feldt condition is as follows: The variances of the differences between any pair of observations on the same EU must be equal; i.e.,

$$Var(y_{ijk} - y_{ij'k}) = 2\lambda \quad \text{for all} \quad j \neq j'$$

Note that compound symmetry implies the Huynh–Feldt condition, but the Huynh–Feldt condition does not imply compound symmetry. A test of the Huynh–Feldt condition, the Mauchly test, is provided in both SAS and SPSS. However, when the sample sizes are relatively small, the Mauchly test has very low power and hence will often fail to detect that the compound symmetry is invalid. This will often result in an incorrect application of the split-plot analysis of a repeated measures experiment.

If the Huynh–Feldt condition is valid, then the split-plot analysis is an appropriate approximation. The model would then be

$$y_{ijk} = \mu + \tau_i + d_{ik} + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$

with $i = 1, \ldots, m$; $j = 1, \ldots, t$; $k = 1, \ldots, n$, where τ_i is the ith treatment effect; β_j is the jth time effect; $(\tau\beta)_{ij}$ is the treatment–time interaction effect; d_{ij}s are independently distributed as $N(0, \sigma_d^2)$ random variables; ε_{ijk}s are indpendently distributed as $N(0, \sigma_\varepsilon^2)$ random variables; and d_{ij} and ε_{ijk} are independently distributed. The above model yields the following variance–covariance structure if the Huynh– Feldt condition is valid: assume $i \neq i', j \neq j', k \neq k'$;

$$Var(y_{ijk}) = \sigma_d^2 + \sigma_\varepsilon^2$$
$$Cov(y_{ijk}, y_{ij'k}) = \sigma_d^2$$

$$Cov(y_{ijk}, y_{i'jk}) = 0$$
$$Cov(y_{ijk}, y_{ijk'}) = 0$$
$$Cov(y_{ijk}, y_{ij'k'}) = 0$$

In the general repeated measures design, measurements from the same EU would likely have a more complex correlation structure, and measurements among EUs in the same treatment group may be correlated. Only measurements from EUs receiving different treatments would be uncorrelated. The condition of compound symmetry yields the following conditions on the variances and covariances of the data:

$$Var(y_{ijk}) = \frac{\sigma_d^2}{1-\rho_d} + \frac{\sigma_\varepsilon^2}{1-\rho_\varepsilon}$$
$$Cov(y_{ijk}, y_{ijk'}) = \frac{\sigma_d^2}{1-\rho_d} + \frac{\sigma_\varepsilon^2\rho_\varepsilon}{1-\rho_\varepsilon}$$
$$Cov(y_{ijk}, y_{ij'k'}) = \frac{\sigma_d^2\rho_d}{1-\rho_d}$$

where ρ_d and ρ_ε are correlation coefficients having values between -1 and $+1$.

An equivalent way to express the above structure on the covariances is given by

$$Var(d_{ij} - d_{ij'}) = 2\sigma_d^2 \quad Var(e_{ijk} - e_{ijk'}) = 2\sigma_\varepsilon^2$$

sphericity condition

The above conditions are called the **sphericity condition.**

The data must be of this form in order for the split-plot analysis to provide an appropriate analysis of the repeated measures experiment.

With $\lambda = t\rho_\varepsilon/2(1-\rho_\varepsilon)$, the AOV table for the split-plot analysis of a repeated measures experiment is given in Table 18.11. In this table, the treatment and time effects are fixed.

Based on Table 18.11, it is clear that the following tests can be performed:

1. H_0: $\theta_{\tau\beta} = 0$
$$F = \frac{\text{MS}_{\text{Trt*Time}}}{\text{MSE}}$$
2. H_0: $\theta_\beta = 0$
$$F = \frac{\text{MS}_{\text{Time}}}{\text{MSE}}$$
3. H_0: $\theta_\tau = 0$
$$F = \frac{\text{MS}_{\text{Trt}}}{\text{MS}_{\text{EU(Trt)}}}$$

TABLE 18.11
Analysis of variance table for a two-factor experiment with repeated measures on one factor

Source	df	Expected Mean Square
TRT	$m-1$	$\sigma_\varepsilon^2(1+2\lambda) + t\sigma_d^2 + nt\theta_\tau$
EU(TRT)	$(n-1)m$	$\sigma_\varepsilon^2(1+2\lambda) + t\sigma_d^2$
Time	$t-1$	$\sigma_\varepsilon^2 + nm\theta_\beta$
TRT*Time	$(m-1)(t-1)$	$\sigma_\varepsilon^2 + n\theta_{\tau\beta}$
Error	$m(t-1)(n-1)$	σ_ε^2
Total	$mtn-1$	

EXAMPLE 18.3

The following example from ***Analysis of Repeated Measures (Crowder and Hand, 1990)*** will be used to illustrate these concepts. In their study, three levels of a vitamin E supplement—zero (control), low, and high—were given to guinea pigs. Five pigs were randomly assigned to each of the three levels of the vitamin E supplement. The weights of the pigs were recorded at 1, 2, 3, 4, 5, and 6 weeks after the beginning of the study (Table 18.12). This is a repeated measures experiment because each pig, the EU, is given only one treatment but each pig is measured six times. The experimenters are interested in the trend in weight over time.

TABLE 18.12
Weight of guinea pigs under three levels of vitamin E

Level of E	Animal	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
C	1	455	460	510	504	436	466
C	2	467	565	610	596	542	587
C	3	445	530	580	597	582	619
C	4	485	542	594	583	611	612
C	5	480	500	550	528	562	576
L	6	514	560	565	524	552	597
L	7	440	480	536	484	567	569
L	8	495	570	569	585	576	677
L	9	520	590	610	637	671	702
L	10	503	555	591	605	649	675
H	11	496	560	622	622	632	670
H	12	498	540	589	557	568	609
H	13	478	510	568	555	576	605
H	14	545	565	580	601	633	649
H	15	472	498	540	524	532	583

a. Plot the weights of the individual pigs versus time, and plot the mean weights versus time for each treatment. Does vitamin E seem to impact the different plots?
b. Test for significant effects on the mean weight of pigs due to level of vitamin E.

Solution

a. The mean weights by level of vitamin E and time are given in Table 18.13.

TABLE 18.13
Weight of guinea pigs under three levels of vitamin E

Level of E	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
C	466.4	519.4	568.8	561.6	546.6	572.0
L	494.4	551.0	574.2	567.0	603.0	644.0
H	497.8	534.6	579.8	571.8	588.2	623.2

The plots of the individual weight gains and a profile plot of the mean weights are given in Figure 18.6 and Figure 18.7, respectively.

FIGURE 18.6
Weight of guinea pigs

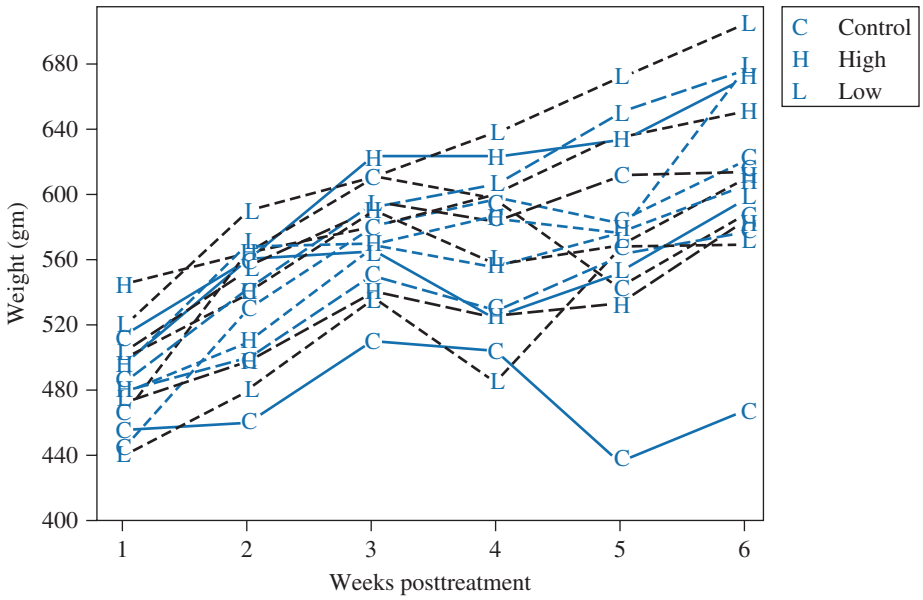
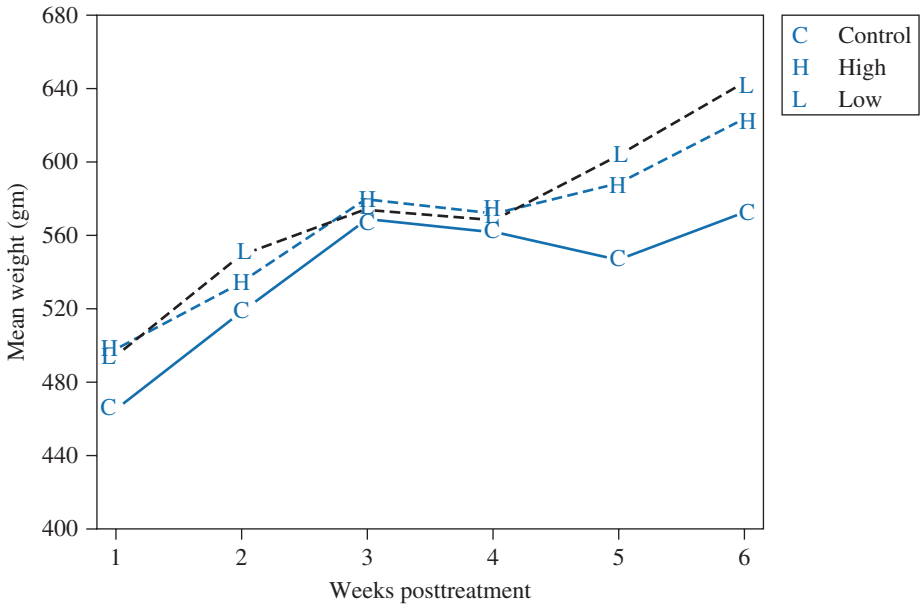


FIGURE 18.7
Mean weight of guinea pigs



b. The following SAS output will be used to obtain the mean squares and F tests.

```
The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source                 DF    Type III SS    Mean Square   F Value   Pr > F

TRT                     2     18548.0667      9274.0333      1.06   0.3782
Error                  12    105434.2000      8786.1833
```

```
The GLM Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

                                                                  Adj Pr > F
Source          DF    Type III SS    Mean Square   F Value   Pr > F    G - G        H - F

WK               5    142554.5000     28510.9000     52.55   <.0001    <.0001       <.0001
WK*TRT          10      9762.7333       976.2733      1.80   0.0801    0.1457       0.1103
Error(WK)       60     32552.6000       542.5433

Greenhouse-Geisser Epsilon  0.4856
Huynh-Feldt Epsilon         0.7191
```

From the above, we obtain the AOV table in Table 18.14.

TABLE 18.14
AOV table for guinea pig experiment

Source	SS	df	MS	*F*	*p*-value
TRT	18,548.07	2	9,274.03	1.06	.3782
PIG (TRT)	105,434.20	12			
Week	142,554.50	5	28,510.90	52.55	$< .0001$
TRT*week	9,762.73	10	976.27	1.80	.0801
Error	32,552.60	60	542.54		

From Table 18.14, we find that there is not significant evidence (p-value = .0801) of an interaction between the treatment and time factors. The profile plot supports this conclusion after taking into account the size of the standard error of the treatment by time sample mean: $\widehat{SE}(\bar{y}_{ij.}) = 19.5780$. Since the interaction was not significant, the main effects of treatment and time can be analyzed separately. The p-value = .3782 for treatment differences, and the p-value $< .0001$ for time differences. The mean weights of the pigs vary across the 6 weeks, but there is not significant evidence of a difference in the mean weights for the three levels of vitamin E supplements. Therefore, the two levels of vitamin E supplement do not appear to provide an increase in the mean weights of the pigs in comparison to the control, which was a zero level of vitamin E supplement. The mean weights appear to follow a cubic relationship with time during the 6 weeks. We could test this conclusion by using contrasts or fitting a regression model to the data.

The above conclusions are all conditional on whether there is significant evidence of a deviation from compound symmetry.

Note that there are three treatments with $r = 5$ replications per treatment for a total of 15 EUs (pigs), each of which is weighed six times for a total of 90 observations. In contrast, a completely randomized design with 90 observations would have 90 EUs, each weighed once. Thus, 75 more pigs are required to perform the completely randomized design. However, this gain in economy has limitations. The inferences are being made about a population of pigs. In the repeated measures design, only 15 pigs from the population are being observed. Thus, there may be greater variability in the estimation of the treatment means due to having such a small sample size per treatment. On the other hand, the repeated measures design allows the researcher to track the behavior of the individual pig over the 6 weeks and hence provides information concerning the potential differences in fluctuations in weight for the individual pigs. The plot of the individual weight data reveals widely varying patterns for the 15 pigs. ■

The F test for the factor treatment is based on between-subject effects and hence is *not* affected by the repeated measures on the factor time. However, the F-ratios for the within-EU effects are affected, and, as with the one-factor experiment with repeated measures, we must worry about the conditions under which these F tests are appropriate. If compound symmetry of the variance–covariance matrix for the y_{ijk}s holds, then we can apply these tests; also, if the Huynh–Feldt conditions alluded to previously hold, then we can apply these F tests. Some (e.g., Greenhouse and Geisser, 1959; Huynh and Feldt, 1970) have suggested that "adjusted" F-values be used to determine the statistical significance of a repeated measures F test when there is some departure from the underlying conditions for that test. The adjustments recommended by the various authors follow the same pattern. A quantity epsilon is defined as a multiplicative adjustment factor for the numerator and denominator degrees of freedom for the F test in question. This epsilon (which we will denote by e) is not to be confused with the random error term ε in our models. For most of these adjustments, the multiplicative factor e ranges between 0 and 1, taking on a value of 1 when the underlying conditions for a valid F test are met and smaller values as the degree of departure from those conditions increases. A value of e having been determined for a given situation, the computed F statistic is compared to the critical value for an F distribution with numerator and denominator degrees of freedom multiplied by e.

The ideas behind the adjustment can be seen if we use the experimental setting for Table 18.11 as the basis for discussion. Here we have a two-factor experiment with repeated measures on the second factor (time). The F tests for the within-EU effects, Time and TRT*time shown in Table 18.11, are valid provided the Huynh–Feldt conditions hold.

For a given experiment, we compute a value of e and adjust the degrees of freedom for the F test by multiplying df_1 and df_2 by e. So to run a test of H_0: $\theta_{\tau\beta} = 0$, a value of e is computed from the sample data. The computed F statistic

$$F = \frac{\mathrm{MS}_{\mathrm{Trt*Time}}}{\mathrm{MSE}}$$

is compared to a critical value, F_α, based on $\mathrm{df}_1 = e(m-1)(t-1)$ and $\mathrm{df}_2 = em(t-1)(n-1)$. Note that when $e = 1$, the underlying conditions hold, and we have the original, recommended degrees of freedom, $\mathrm{df}_1 = (m-1)(t-1)$ and $\mathrm{df}_2 = m\,(t-1)(n-1)$.

In experimental situations where repeated measures data are to be analyzed and where you have access to SAS, you can use PROC GLM to compute revised p-values for two different adjustments to the degrees of freedom. The first adjustment, proposed by Greenhouse and Geisser (1959), uses a sample estimate of e. This adjustment, labeled "G–G" in the SAS output, has been shown, in simulation studies, to be ultraconservative because the actual p-value may be much smaller than that indicated by the p-value using the G–G adjustment. The second adjustment factor (proposed by Huynh and Feldt, 1970) is based on a different formula for e. Once again, however, an estimate of this adjustment factor is computed from the sample data. The degrees of freedom for critical values of the F statistics are then adjusted using the estimate of e. This adjustment is labeled "H–F" in the PROC GLM output. Although the Greenhouse–Geisser e and Huynh–Feldt e both must be in the interval $0 < e \leq 1$, the H–F estimate of e can sometimes be greater than 1. In these situations, a value of $e = 1$ is used in determining the appropriate degrees of freedom for the F test.

EXAMPLE 18.4

Refer to the SAS output for Example 18.3.

a. Locate the estimated values for the Greenhouse–Geisser adjustment factor and the Huynh–Feldt adjustment factor.
b. Are the conclusions for the tests on time effects and the time–vitamin E interaction affected by these adjustments?

Solution

a. The Greenhouse–Geisser estimate of e is .4856, and the Huynh–Feldt estimate of e is .7191.
b. Time Effects: F tests based on the G–G adjustment and on the H–F adjustment yield p-values of $<.0001$ and $<.0001$, respectively, which are the same as the values from the original F test. The adjustments did not change the conclusion obtained from the unadjusted F test.

Time by Treatment Interaction Effects: F tests based on the G–G adjustment and on the H–F adjustment yield p-values of .1457 and .1103, respectively. These values are somewhat higher than the p-value from the original F test, .0801. The adjustments would not change the conclusion obtained from the unadjusted F test if an $\alpha = .05$ value was used but would change the conclusion if a higher type I error rate was used, such as $\alpha = .10$. For $\alpha = .10$, the unadjusted F test would have declared the interaction effect significant, whereas the G–G and H–F adjusted F tests would not. ■

18.5 Crossover Designs

We will now consider an extension to the single-factor experiment discussed in Section 18.3. Recall that in Table 18.7 we presented data for an experimental situation in which each of the n patients received the same three treatments in a random order. Thus, each patient was observed n times in the experiment. It is important to emphasize the difference between a crossover design and the general repeated measures design. In a repeated measures experiment, the experimental unit receives a treatment and then the experimental unit has multiple observations or measurements made on it over time or space. The experimental unit does not receive a new treatment, between successive measurements.

In a crossover design, each experimental unit is observed under each of the t treatments during t observation times. That is, every experimental unit has multiple treatments applied to it, and then a new measurement or observation is obtained. Because the treatments are compared on the same experimental units, the between-experimental unit variation is greatly reduced. The individual experimental units serve as blocks in order to reduce the experimental variation (reduced SSE) and hence there is an increase in the efficiency of the estimation of the treatment means.

When comparing treatments, the effect of the time period in which the treatment was applied comes into the analysis. Differences in observations may be due to treatment differences and/or time period differences. Crossover designs are constructed to avoid confounding the time period effects with the treatment effects.

EXAMPLE 18.5

Suppose we have three treatments—T_1, T_2, and T_3–with each treatment applied to each of 12 patients during three time periods—P_1, P_2, and P_3. The drugs were applied in the same order to all 12 patients, as shown in Table 18.15.

TABLE 18.15
Design layout for Example 18.5

	Time Period				Time Period		
Patients	**1**	**2**	**3**	**Patients**	**1**	**2**	**3**
1	T_1	T_2	T_3	7	T_1	T_2	T_3
2	T_1	T_2	T_3	8	T_1	T_2	T_3
3	T_1	T_2	T_3	9	T_1	T_2	T_3
4	T_1	T_2	T_3	10	T_1	T_2	T_3
5	T_1	T_2	T_3	11	T_1	T_2	T_3
6	T_1	T_2	T_3	12	T_1	T_2	T_3

Suppose that from the data collected under the above design a large difference was observed in the treatment means: $\bar{y}_{1..}$, $\bar{y}_{2..}$, and $\bar{y}_{3..}$. Was this difference due to treatment differences or time period differences?

Solution With the above design, it would be impossible to determine. The sample mean responses for estimating the effects of the three treatment means are identical to the sample mean responses for estimating the effects of the three time period means. That is, with this design, the effects of treatment and time period are confounded.

To avoid the confounding of the treatment and time period effects, it is necessary to consider multiple sequences in which the treatments are administered to the experimental units. There are $3! = 6$ possible sequences in which the three treatments could be administered to the 12 subjects during the three treatment periods. Table 18.16 lists those sequences.

TABLE 18.16
Sequences for administrating three treatments in three time periods

	Time Period		
Sequence	**1**	**2**	**3**
1	T_1	T_2	T_3
2	T_2	T_3	T_1
3	T_3	T_1	T_2
4	T_2	T_1	T_3
5	T_3	T_2	T_1
6	T_1	T_3	T_2

The experimenter could randomly assign two patients to each of the six sequences. This would eliminate the confounding among the effects due to treatments, sequences, and time periods. Every treatment would be observed in every sequence and in every time period. In many experiments, the researcher will select a subset of all $t!$ sequences in order to increase the number of subjects per sequence. This yields a more accurate assessment of the sequence effect. ■

EXAMPLE 18.6

Twelve males volunteered to participate in a study to compare the effect of three formulations of a drug product: formulation 1 was a 5-mg tablet, formulation 2 was a 100-mg tablet, and formulation 3 was a sustained-release capsule. Suppose it is decided to use only three of the six sequences listed in Table 18.16. Select three of the six possible sequences, and describe how to randomize this experiment. Also, include a model for this experiment.

Solution The experimenter selected the first three of the six sequences and randomly assigned four subjects to each sequence. On each treatment day, volunteers were given their assigned formulation and were observed to determine the duration of effect of the treatment (blood pressure lowering). The data would be as shown in Tables 18.17 and 18.18.

TABLE 18.17
Design layout for Example 18.6

	Time Period		
Sequence	**1**	**2**	**3**
1	T_1	T_2	T_3
2	T_2	T_3	T_1
3	T_3	T_1	T_2

TABLE 18.18
Blood pressure data

		Time Period		
Sequence	**Patient (Seq)**	**1**	**2**	**3**
1	1	1.5	2.2	3.4
	2	2.0	2.6	3.1
	3	1.6	2.7	3.2
	4	1.1	2.3	2.9
2	1	2.5	3.5	1.9
	2	2.8	3.1	1.5
	3	2.7	2.9	2.4
	4	2.4	2.6	2.3
3	1	3.3	1.9	2.7
	2	3.1	1.6	2.5
	3	3.6	2.3	2.2
	4	3.0	2.5	2.0

A model for this experiment would be the following. Let y_{ijk} be the response observed in time period k from the jth patient in sequence i.

$$y_{ijk} = \mu + \delta_i + \beta_{j(i)} + \gamma_k + \tau_{d(i,k)} + \varepsilon_{ijk}$$

with δ_i, $i = 1, 2, 3$, as the fixed sequence effect; $\beta_{j(i)}$, $j = 1, 2, 3, 4$, as the random patient within sequence effect; γ_k, $k = 1, 2, 3$, as the fixed time period effect; $\tau_{d(i,k)}$, $d = 1, 2, 3$, as the fixed treatment effect; and ε_{ijk} as the random experimental error effect. ■

The general setting of a crossover design will now be described. Suppose we have t treatments that are to be compared with respect to their mean responses. In the experiment, we have either very heterogeneous experimental units or a limited number of experimental units and decide that each experimental unit will be observed under all t treatments. The experimental units serve as blocks and thus control the variation in response from experimental unit to experimental unit for a given treatment. An obvious question of concern is whether or not the order in which the experimental unit receives the treatments has an effect on the responses. There are $t!$ possible sequences in which the t treatments may be applied. Generally only a subset of the $t!$ possible sequences will be used in the study. The experimenter decides on n sequences that are of greatest interest. There will be

r_i experimental units randomly assigned to the ith treatment sequence, which will be observed during p time periods. There is generally a time delay between when the treatment is administered and when the response is measured on the experimental unit. Furthermore, after the measurement is taken, there will be a further delay before the next treatment is applied in order that the previously administered treatment will not have a **carryover effect** on the experimental unit during the administering of the next treatment. This is called the **washout period.** The following model would be applicable:

carryover effect
washout period

$$y_{ijkdc} = \mu + \delta_i + \beta_{j(i)} + \gamma_k + \tau_{d(i,k)} + \lambda_{c(i,k)} + \varepsilon_{ijk}$$

with μ the overall mean response; δ_i, $i = 1, \ldots, n$, the fixed effect of the ith sequence; $\beta_{j(i)}$, $j = 1, \ldots, r_i$, the random effect for the jth experimental unit within the ith sequence; γ_k, $k = 1, \ldots, p$, the kth fixed time period effect; $\tau_{d(i,k)}$ the direct effect of the treatment applied during period k in sequence i; and $\lambda_{c(i,k)}$ the carryover effect of the treatment applied during period k in sequence i.

Note that there is randomization of the subjects to the sequences. Furthermore, there are two sizes of experimental units. The **experimental unit for sequence** is **"subject,"** and the **experimental unit for treatment** is **"time period."** The sequence effect measures some form of the time period by treatment interaction and may be an indication of a carryover effect and/or correlation in the measurements over time periods.

The analysis of variance table for a three-period crossover design with three sequences (fixed effects), n subjects per sequence (random effect), three treatments (fixed effects), three time periods (fixed effects), and a fixed carryover effect is given in Table 18.19.

In those studies in which the carryover effect is found to be highly significant, the tests for treatment effects would be confounded with the carryover effects. This would invalidate the conclusions about the treatment differences due to the fact that the order in which the treatments were applied to the subjects has a significant effect on the responses. In the case that the carryover effect is significant, the overall conclusions about the treatment effects would be in question. However, there is still information in the study that can be used in assessing treatment effects. The data from the **first time period** can be used in testing for treatment effects because there would be no carryover from any previous applications of the treatments.

first time period

A particularly unique characteristic of the crossover design is that each subject receives all t treatments. A degree of balance is obtained in the crossover design by having each treatment follow every other treatment the same number of times in the study, having each treatment occur the same number of times in each time period, and observing each treatment only once on each experimental unit. These characteristics create some particular advantages and disadvantages for the crossover design.

TABLE 18.19
Analysis of variance for a crossover design

Source	df	Expected Mean Square
Sequence	2	$\sigma_\varepsilon^2 + 3\sigma_\beta^2 + 3n\theta_\delta$
Patient (Seq)	$3(n-1)$	$\sigma_\varepsilon^2 + 3\sigma_\beta^2$
Period	2	$\sigma_\varepsilon^2 + 3n\theta_\gamma$
Treatment	2	$\sigma_\varepsilon^2 + 3n\theta_\tau$
Carryover	2	$\sigma_\varepsilon^2 + 3n\theta_\lambda$
Error	$3(2)(n-1)$	σ_ε^2
Total	$9n-1$	

Advantages:

1. Reduction in the between-experimental unit variation (subject is serving as a blocking variable)
2. Increased precision in comparing treatment means
3. Reduction in the experimental cost when experimental units are expensive and/or difficult to recruit for study and/or difficult or expensive to maintain during study

Disadvantages:

1. May be a carryover effect, which will invalidate much of the study
2. Reduced information about and coverage of the population of experimental units

There is a further complication with the above model besides the potential of the carryover effect. There are t observations on each experimental unit under the t different treatments. Thus, we have a multivariate response on each experimental unit, not a single response. Under special conditions, which were discussed in the repeated measures section of this text, we can validly analyze the data as a univariate experiment. Furthermore, if there was not a carryover effect, then we could analyze the experiment as a Latin Square design with blocking variables sequence and time period. In order to test for the carryover effect in the model, it is necessary to create a new variable to be included in the data analysis. The carryover variable is defined as follows:

1. Let C_{ijk} be the value of the carryover variable for the jth experimental unit in the ith sequence during the kth period.
2. All values of C_{ikj} are set equal to 0 during period 1: $C_{ij1} = 0$ for all ij.
3. The values of C_{ikj} are values for the treatment variable in period $k - 1$.

We will illustrate these ideas in the following example.

EXAMPLE 18.7

Refer to the experimental data in Example 18.6. Using the data from Example 18.6, construct the carryover variable necessary for testing for a carryover effect. Then conduct an analysis of variance and test for carryover and direct treatment effects.

Solution Using the notation S = sequence, EU = patient, T = treatment, P = period, and CAR = carryover, we obtain the data shown in Table 18.20.

TABLE 18.20
Data structure for evaluating carryover effect

S	EU	T	P	y	CAR	S	EU	T	P	y	CAR	S	EU	T	P	y	CAR
1	1	T_1	1	1.5	0	1	1	T_2	2	2.2	T_1	1	1	T_3	3	3.4	T_2
1	2	T_1	1	2.0	0	1	2	T_2	2	2.6	T_1	1	2	T_3	3	3.1	T_2
1	3	T_1	1	1.6	0	1	3	T_2	2	2.7	T_1	1	3	T_3	3	3.2	T_2
1	4	T_1	1	1.1	0	1	4	T_2	2	2.3	T_1	1	4	T_3	3	2.9	T_2
2	1	T_2	1	2.5	0	2	1	T_3	2	3.5	T_2	2	1	T_1	3	1.9	T_3
2	2	T_2	1	2.8	0	2	2	T_3	2	3.1	T_2	2	2	T_1	3	1.5	T_3
2	3	T_2	1	2.7	0	2	3	T_3	2	2.9	T_2	2	3	T_1	3	2.4	T_3
2	4	T_2	1	2.4	0	2	4	T_3	2	2.6	T_2	2	4	T_1	3	2.3	T_3
3	1	T_3	1	3.3	0	3	1	T_1	2	1.9	T_3	3	1	T_2	3	2.7	T_1
3	2	T_3	1	3.1	0	3	2	T_1	2	1.6	T_3	3	2	T_2	3	2.5	T_1
3	3	T_3	1	3.6	0	3	3	T_1	2	2.3	T_3	3	3	T_2	3	2.2	T_1
3	4	T_3	1	3.0	0	3	4	T_1	2	2.5	T_3	3	4	T_2	3	2.0	T_1

Note that the carryover effect variable, CAR, has all zeros in period 1. The values of CAR in period 2 are identical for the values of TRT in period 1, and the values of CAR in period 3 are identical to the values of TRT in period 2. The following output from SAS will provide us with the appropriate tests of the carryover effect.

```
CROSSOVER DESIGN WITH TEST FOR CARRYOVER
MODEL WITH BOTH TRT AND CARRYOVER

The GLM Procedure

Dependent Variable: Y

                                 Sum of
Source                    DF       Squares   Mean Square   F Value   Pr > F
Model                     15   10.43750000    0.69583333      5.13   0.0005
Error                     20    2.71222222    0.13561111
Corrected Total           35   13.14972222

Source                    DF   Type III SS   Mean Square   F Value   Pr > F
Seq                        2    0.23388889    0.11694444      0.86   0.4373
Pat(Seq)                   9    0.66916667    0.07435185      0.55   0.8221
Trt                        2    9.51722222    4.75861111     35.09   <.0001
Per                        2    0.01722222    0.00861111      0.06   0.9387

Dependent Variable: Y

                                 Sum of
Source                    DF       Squares   Mean Square   F Value   Pr > F
Model                     17   11.08638889    0.65214052      5.69   0.0003
Error                     18    2.06333333    0.11462963
Corrected Total           35   13.14972222

Source                    DF   Type III SS   Mean Square   F Value   Pr > F

Seq                        2    0.05583333    0.02791667      0.24   0.7864
P(Seq)                     9    0.66916667    0.07435185      0.65   0.7425
Trt                        2    3.98433333    1.99216667     17.38   <.0001
Period                     1    0.00041667    0.00041667      0.00   0.9526
Carry                      2    0.64888889    0.32444444      2.83   0.0853

Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: Y

Source                    DF   Type III SS   Mean Square   F Value   Pr > F
Seq                        2      0.055833      0.027917      0.30   0.7466


Error                 26.568   2.510443      0.094491
Error: 0.5*MS(P(Seq)) + 0.5*MS(Error)
```

In the above output, it was necessary to run two models, one with the carryover effect and one without the carryover effect, in order to obtain the sum of squares for period. We will summarize the information from the SAS output into the AOV table in Table 18.21, in which sequence, time period, direct effect of formulations, and carryover are fixed effects and patient in sequence is a random effect.

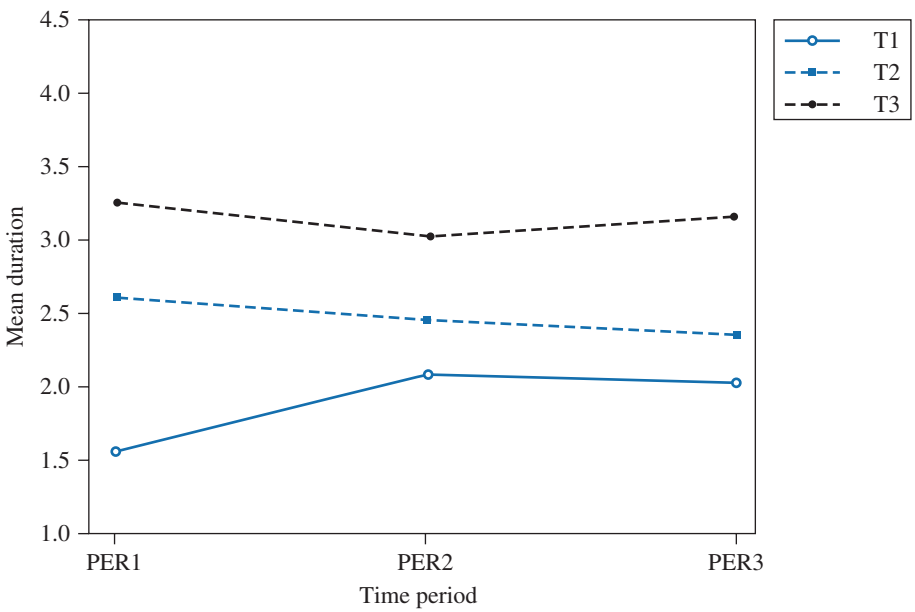
First, we examine the carryover effect. The p-value from the F test is .0853. Thus, there is a hint of a carryover effect, but it is not significant at the .05 level. The carryover effect is imbedded in the time period by treatment interaction. Figure 18.8 is a plot of the treatment means (mean duration for each formulation) by time period. This reveals an indication of an interaction between time period and treatment. Although formulation 3 has the highest mean duration followed by

TABLE 18.21
AOV table for evaluating carryover effect

Source	df	Sum of Squares	Mean Square	*F*	*p*-value
Seq	2	.0558	.0279	.30	.7466
P(Seq)	9	.6692	.0744		
Trt	2	3.9843	1.9922	17.38	<.0001
Period	2	.0172	.0086	.06	.9387
Carry	2	.6489	.3244	2.83	.0853
Error	18	2.0633	.1146		
Total	35	13.1497			

formulation 2 and then formulation 1 in all three time periods, the amount of difference in the three formulations is considerably more in period 1 than in the other two time periods. However, after taking into account the variability in the treatment means, the interaction is found to be nonsignificant. Therefore, we can next examine the direct effect of the treatment: drug formulations. The *F* test for a direct effect of formulations on mean duration is highly significant (p-value $< .0001$). A Tukey multiple-comparison analysis of the three formulations reveals that all pairs of treatment means are significantly different at the .05 level.

FIGURE 18.8
Profile plot of mean duration versus time period for three formulations



When there are only two compounds to be examined, the Latin square arrangement, called a two-period crossover design, would have $2n$ patients randomly assigned to the two sequences, n to each sequence. The two-period crossover design is shown in Table 18.22.

The model for this experiment is

$$y_{ijkl} = \mu + \delta_i + \beta_{j(i)} + \gamma_k + \tau_l + \varepsilon_{ijkl}$$

where δ_i is the fixed effect due to sequence i, $\beta_{j(i)}$ is the random patient j in sequence i effect, γ_k is the fixed time period effect, τ_l is the fixed effect due to treatment l, and ε_{ijkl} is the random experimental error effect.

TABLE 18.22
Layout for a two-period crossover design

		Factor B (periods)	
Sequence	**Patient**	**1**	**2**
1	n	A_1	A_2
2	n	A_2	A_1

TABLE 18.23
AOV table for a two-period crossover design

Source	**SS**	**df**	**EMS (A, B fixed; patient random)**
Sequence	SSSeq	1	$\sigma_\varepsilon^2 + 2\sigma_\beta^2 + 2n\theta_\delta$
Patient(Seq)	SSP(Seq)	$2(n-1)$	$\sigma_\varepsilon^2 + 2\sigma_\beta^2$
Treatment	SSA	1	$\sigma_\varepsilon^2 + 2n\theta_\tau$
Period	SSB	1	$\sigma_\varepsilon^2 + 2n\theta_\gamma$
Error	SSE	$2(n-1)$	σ_ε^2
Totals	TSS	$4n-1$	

Note there is no carryover term in this model. We must assume this term is negligible; otherwise, the design is inappropriate because there are no degrees of freedom available for testing the significance of the carryover effect. The AOV table for a two-period crossover design is shown in Table 18.23.

There are many other extensions to the repeated measures designs discussed in this chapter. For example, one could combine the concept of repeated measures on the same factor illustrated in Table 18.7 with the crossover design. Such a plan is illustrated in Table 18.24. Thus, rather than taking one observation per patient within each period, we would take observations at t different time points. For example, we could measure blood pressure every 15 minutes for the first hour following treatment with compound i and then hourly for the next 7 hours. This would be done in each of the periods for a total of 10 blood pressure measurements on each patient in each time period.

Although we will not give the analysis of variance for this extension to the repeated measures experiments discussed in this chapter and will not cover other, more complicated repeated measures designs, we want you to be aware of the wealth of possible designs that are available if you are willing to take more than one observation per experimental unit. The interested reader is referred to Vonesh and Chinchilli (1997); Crowder and Hand (1990); Jones and Kenward (2015); and Diggle, Liang, and Zeger (1996).

TABLE 18.24
Two-period crossover design with repeated measures

	Period	
	1	**2**
	Time	**Time**
Sequence	**1 2 . . . t**	**1 2 . . . t**
1	A_1	A_2
2	A_2	A_1

18.6 RESEARCH STUDY: Effects of an Oil Spill on Plant Growth

On January 7, 1992, an underground oil pipeline ruptured and caused the contamination of a marsh along the Chiltipin Creek in San Patricio County, Texas. The cleanup process consisted of burning the contaminated regions in the marsh. To evaluate the influence of the oil spill on the flora, the researchers designed a study of plant growth after the burning was completed. They focused their findings on *Distichlis spicata,* a flora of particular importance to the area of the spill. Two questions of importance to the researchers were as follows:

1. Did the oil site recover after the spill and burning?
2. How long did it take for the recovery?

To answer these questions, the researchers needed to have a baseline to which they could compare the *Distichlis spicata* density in the months after the burning of the site. The density of the flora depended on soil characteristics, slope of the land, environmental conditions, weather, and many other factors. The researchers designated as the control site a nearby section of land that was not affected by the oil spill but that had soil and environmental properties similar to those of the spill site. At both the oil spill site and the control site, 20 tracts were randomly chosen. After a 9-month transition period, measurements were taken at approximately 3-month intervals for a total of eight time periods. During each time period, the number of *Distichlis spicata* within each of the 40 tracts was recorded.

The experimental design is a repeated measures design with two treatments, the oil spill and the control region, and eight measurements taken over time on each of the tracts over a 2-year period. The data consisted of the number of *Distichlis spicata* plants found on each tract during the eight observation periods on both the control and the burned (oil spill) sites. There were a total of 320 data values, as displayed in Table 18.2. The mean flora counts by treatment and date are given in Table 18.25.

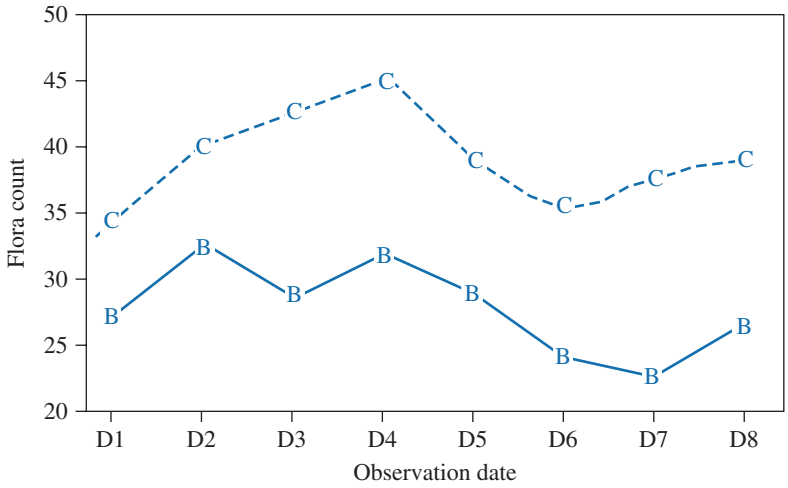
Analyzing the Data

The flora counts were plotted in Figure 18.1, using boxplots for each date and treatment. The boxplots reveal that the control plots have higher median flora counts than the oil spill plots. The control plots, however, are somewhat more variable than the oil spill plots. This may be due to the burning treatment that was used on the oil spill plots, which often results in more homogeneous tract conditions than those that were present prior to the burning. The objective of the study was to examine the effects of the oil spill and subsequent burning of the tracts on which the oil spill occurred on the density of the flora *Distichlis spicata.* Since baseline density of the flora prior to the oil spill and burning did not exist, a comparison will be made with tracts that were not involved in the oil spill. In Figure 18.9, a profile

TABLE 18.25
Flora count means by treatment and date

	Inspection Date							
Treatment	**Oct-92**	**Jul-93**	**Oct-93**	**Jan-94**	**Apr-94**	**Jul-94**	**Oct-94**	**Jan-95**
Burned	27.20	32.65	28.60	31.95	28.90	24.10	22.55	26.65
Control	34.45	40.25	42.80	45.10	39.05	35.55	37.75	39.35

FIGURE 18.9
Profile plot of flora densities by type and date



plot of the flora densities is displayed for the control and burned tracts across the eight observation dates. The mean densities for the control (C) tracts are consistently higher than the mean densities for the burned (B) tracts. The changes in mean densities have similar trends except on two of the observation dates (D3 and D7). On these two dates, the mean density of the flora on the burned tracts had a decrease from the previous date, whereas the mean density for the control plots increased. We will next construct the repeated measures AOV to confirm these observations.

An analysis of the data yields the AOV table in Table 18.26 for the flora density data.

There is a highly significant date by treatment interaction, which confirms the observations we had made from examining the profile plot. Furthermore, there is a significant difference between the mean densities of the burned and control plots. The control plots had larger mean flora densities than the burned plots. This difference was 7.25 at the first observation date and increased to a final difference of 12.70 on the final observation date, slightly more than 2 years later. The tracts on which the oil spill occurred showed no recovery in mean flora density, dropping from 27.20 on October 1992 to 26.65 on January 1995. Since the flora density on the control tracts, which had similar soil conditions and environmental exposures during the study period, increased from 34.45 to 39.35, we would conclude that the oil spill and subsequent burning resulted in reduced flora density on the affected tracts.

TABLE 18.26
AOV table for research study

Source	SS	df	MS	*F*	*p*-value	Adj G–G	*p*-value H–F
Treatment	10,511.11	1	10,511.11	6.56	.0045		
Tracts in treatment	60,844.63	38	1,601.17				
Date	2,845.09	7	406.44	19.35	.0001	.0001	.0001
Date × treatment	602.29	7	86.04	4.10	.0001	.0001	.0001
Error	5,587.88	266	21.01				

Greenhouse–Geisser Epsilon = .5269
Huynh–Feldt Epsilon = .5355

18.7 Summary

In this chapter, we have discussed some of the initial concepts and designs associated with split-plot and repeated measures experiments. We introduced single- and two-factor experiments, analyses for these experiments, and the special topics of two- and three-period crossover designs. These methods are only a beginning, however. Rather than presenting an exhaustive, detailed account of the subject, we have looked at these few situations to see the applicability and utility of some of the repeated measures designs and procedures. Facility in designing and analyzing such experiments can be gained only after more detailed study of repeated measures topics through additional reading and course work.

18.8 Exercises

18.2 Split-Plot Designed Experiments

Basic

18.1 An experiment is to be designed as a completely randomized design with a split-plot treatment assignment. Suppose the wholeplot treatment (A) has four levels and the split-plot treatment (B) has three levels. There are a total of 10 replications of the wholeplot treatment. Assume that both factors A and B have fixed levels.

- **a.** Describe a method of randomizing the experimental units to the levels of factors A and B in this experiment.
- **b.** Write a linear model for this experiment. Make sure to identify each of the terms in the model and list the range of values for all subscripts.
- **c.** Construct an analysis of variance table for this experiment, including columns for sources of variation, degrees of freedom, and expected mean squares.

Basic

18.2 An experiment is to be designed as a completely randomized design with a split-plot treatment assignment. Suppose the wholeplot treatment A has four levels. The split-plot treatments consist of the cross of two factors: factor B having three levels and factor C with two levels. There are a total of five replications of the wholeplot treatment and three of the split treatments. Assume that factors A, B, and C have fixed levels.

- **a.** Describe a method of randomizing the experimental units to the levels of factors A, B, and C in this experiment.
- **b.** Write a linear model for this experiment. Make sure to identify each of the terms in the model and list the range of values for all subscripts.
- **c.** Construct an analysis of variance table for this experiment, including columns for sources of variation, degrees of freedom, and expected mean squares.

Basic

18.3 An experiment is to be designed as a randomized complete block experiment with three blocks and a split-plot treatment assignment. Each block is divided into four units. Suppose the wholeplot treatment (A) has four levels and the split-plot treatment (B) has three levels with both factors having fixed levels. The researcher wants to have two replications of each level of factor B in each block.

- **a.** Describe a method of randomizing the experimental units to the levels of factors A and B in this experiment.
- **b.** Write a linear model for this experiment. Make sure to identify each of the terms in the model and list the range of values for all subscripts.
- **c.** Construct an analysis of variance table for this experiment, including columns for sources of variation, degrees of freedom, and expected mean squares.

Sci.

18.4 A meat science researcher designed a study to investigate the impact of increasing the portion of grain (and hence decreasing the portion of hay) in the daily ration for cattle on the tenderness of beef steaks obtained from the cattle. Twelve steers of the same breed, age, and weight were selected for the study. Four of the steers were randomly assigned to one of the following three rations, factor A:

Ration 1: (A_1) 75% grain, 25% hay

Ration 2: (A_2) 50% grain, 50% hay

Ration 3: (A_3) 25% grain, 75% hay

After being on the ration for 90 days, the steers were butchered, and four sirloin steaks were obtained from each carcass. The steaks were then randomly assigned to one of four aging times, factor B: 1, 7, 14, or 21 days. After being stored at 1°C for 90 days, the steaks were thawed and then cooked to an internal temperature of 70°C. Next, 68 cores (1.27 cm in diameter) were removed parallel to fiber orientation from each steak, and the peak shear force was measured on each core using a Warner–Bratzler shearing device. The mean shear force values (kg) are given in the following table.

	Ration 1				Ration 2				Ration 3				
Age	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Mean**
1	3.1	3.2	4.9	6.0	4.9	5.9	3.1	4.6	5.3	4.8	4.7	4.9	4.62
7	2.9	2.1	4.1	5.2	5.2	5.8	2.8	4.4	5.2	4.6	4.5	4.8	4.30
14	2.4	2.5	3.4	5.1	4.4	5.1	3.1	4.7	5.1	4.2	4.2	4.7	4.08
21	2.1	2.1	3.7	5.0	4.6	4.9	2.1	3.8	4.8	4.1	3.8	4.4	3.78
Mean		3.61				4.34				4.63			

The treatment means are given in the following table.

	Age				
Ration	**1**	**7**	**14**	**21**	**Mean**
1	4.30	3.58	3.35	3.23	3.61
2	4.63	4.55	4.33	3.85	4.34
3	4.93	4.78	4.55	4.28	4.63
Mean	4.62	4.30	4.08	3.78	

a. Provide the linear model for this study. Include the ranges on all subscripts.
b. Provide a profile plot that will allow an assessment of the age by ration interaction.
c. Based on the table of means and your profile plot, does the decrease in mean shear force with increased aging of the steaks appear to be the same for all three rations?

18.5 Refer to Exercise 18.4.
a. Construct an analysis of variance table for this study.
b. Is there a significant interaction between age and type of ration?
c. Are there significant differences in the mean shear forces for the three rations?
d. Are there significant differences in the mean shear forces for the four aging times?

18.6 Refer to Exercise 18.4.
a. Explain how this study could have been conducted as a completely randomized design.
b. What would be the gain in conducting the experiment as a completely randomized design over the split-plot design?
c. If the completely randomized design is an improvement over the split-plot design, why was the split-plot design used?

18.4 Two-Factor Experiments with Repeated Measures on One of the Factors

Env.

18.7 The cayenne tick is recognized as a pest of wildlife, livestock, and humans. It is distributed in the Western Hemisphere between 30°N and 30°S latitude. This tick has been identified as a potential vector of several diseases, but the ecology of the cayenne tick is poorly understood. The following study was conducted to examine the survival potential of this tick as a function of the saturation deficit (SD) of the environment. Saturation deficit is an index of environmental

conditions that combines both temperature and relative humidity, with SD increasing with temperature but decreasing with relative humidity. Thus, high values of SD are associated with high temperatures and low relative humidities, conditions that cause ticks to experience maximum water loss. Five values were selected for SD (2.98, 4.83, 5.80, 8.88, and 13.38 mmHg) for use in the study. The conditions were established in an artificial environment, with five ticks randomly assigned to each of these conditions. The whole-body water loss of the ticks was recorded every 2 days over approximately a 3-week study period. The water losses (mg) of the ticks are given here.

		Days of Exposure										
SD	**Tick**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
2.98	1	.54	.59	.64	.73	.76	.89	.93	1.01	1.08	1.15	1.23
	2	.69	.75	.81	.90	.97	1.20	1.14	1.19	1.26	1.38	1.43
	3	.77	.80	.87	.94	1.01	1.10	1.17	1.24	1.34	1.41	1.51
	4	.64	.69	.77	.83	.88	.96	1.04	1.09	1.20	1.23	1.31
	5	.51	.58	.62	.71	.74	.81	.88	.93	.99	1.03	1.13
4.83	1	.64	.71	.77	.89	.90	1.00	1.06	1.14	1.22	1.34	1.39
	2	.80	.91	.97	1.01	1.11	1.19	1.29	1.31	1.37	1.47	1.54
	3	.79	.85	.89	.99	1.04	1.05	1.16	1.21	1.32	1.39	1.47
	4	.77	.82	.88	.92	1.01	1.09	1.19	1.27	1.35	1.44	1.58
	5	.79	.84	.91	.98	1.07	1.14	1.19	1.31	1.37	1.46	1.55
5.80	1	.72	.79	.83	.94	.98	1.09	1.12	1.21	1.28	1.34	1.41
	2	.89	.94	1.01	1.21	1.27	1.40	1.44	1.49	1.49	1.58	1.63
	3	.97	.99	1.07	1.09	1.21	1.30	1.37	1.44	1.54	1.61	1.73
	4	.85	.88	.97	1.05	1.09	1.17	1.24	1.29	1.30	1.23	1.51
	5	.71	.78	.82	.91	.94	1.11	1.19	1.23	1.29	1.33	1.43
8.88	1	.93	.99	1.03	1.14	1.18	1.29	1.33	1.36	1.38	1.54	1.62
	2	1.09	1.14	1.21	1.41	1.47	1.55	1.64	1.69	1.71	1.78	1.83
	3	1.19	1.20	1.07	1.29	1.31	1.50	1.57	1.64	1.74	1.81	1.93
	4	1.05	1.08	1.17	1.25	1.29	1.37	1.44	1.49	1.50	1.53	1.71
	5	1.01	1.09	1.18	1.21	1.29	1.31	1.39	1.43	1.49	1.53	1.63
13.38	1	1.05	1.09	1.13	1.24	1.28	1.39	1.43	1.56	1.68	1.74	1.82
	2	1.29	1.34	1.41	1.51	1.57	1.65	1.74	1.79	1.83	1.88	1.93
	3	1.38	1.40	1.47	1.49	1.51	1.60	1.69	1.74	1.79	1.87	2.03
	4	1.23	1.28	1.37	1.45	1.49	1.57	1.64	1.69	1.70	1.73	1.81
	5	1.23	1.29	1.38	1.41	1.49	1.52	1.48	1.53	1.59	1.63	1.78

a. Display the profile plot for these data, showing mean whole-body weight loss by time period for each value of SD.
b. Does an increase in SD appear to increase the whole-body weight loss for the cayenne tick?

18.8 Refer to the data in Exercise 18.7.
a. Provide a model for this design.
b. Construct an AOV table for the study.
c. Is there significant evidence that an increase in SD results in an increase the whole-body weight loss for the cayenne tick? Use $\alpha = .05$.
d. Is the increase in whole-body weight loss for the cayenne tick over the study the same for all levels of SD? Use $\alpha = .05$.

Med. **18.9** An antihistamine is frequently studied using a model to examine its effectiveness (compared to a placebo) in inhibiting a positive skin reaction to a known allergen. Consider the following situation. Individuals are screened to find 20 subjects who demonstrate sensitivity to the allergen to be used in the study. The 20 subjects are then randomly assigned to one of two treatment groups

(the known antihistamine and an identical-appearing placebo), with 10 subjects per group. At the start of the study, a baseline (predrug) sensitivity reading is obtained, and then each patient begins taking the assigned medication for 3 days. Skin sensitivity readings are taken at 1, 2, 3, 4, and 8 hours following the first dose. The percentage inhibition of skin sensitivity reaction (reduction in swelling of the area where the allergen is applied compared to the baseline) is shown here for each of the 20 patients.

		Time (hours)				
Treatment	**Patient**	**1**	**2**	**3**	**4**	**8**
Antihistamine	1	10.5	28.2	15.3	43.0	29.0
	2	41.2	25.3	27.8	28.0	53.2
	3	43.0	20.8	29.3	5.2	26.5
	4	61.4	61.6	62.8	43.8	19.6
	5	5.0	28.2	31.6	19.5	2.3
	6	−10.2	27.2	38.1	35.5	18.0
	7	−12.9	22.1	34.0	43.4	34.2
	8	27.1	26.5	38.8	28.5	17.4
	9	13.0	19.7	23.5	29.4	39.6
	10	28.9	26.1	11.2	18.1	16.5
Placebo	1	3.0	9.3	1.0	15.0	3.0
	2	−1.5	−10.1	20.2	18.3	13.5
	3	10.8	20.6	28.3	25.2	15.8
	4	15.3	19.8	25.4	31.3	21.7
	5	8.7	8.0	17.5	26.6	16.4
	6	−4.6	5.8	12.7	15.6	29.6
	7	−16.6	28.4	32.7	34.4	15.8
	8	9.4	15.7	22.7	29.8	23.2
	9	−19.3	15.7	21.7	30.4	26.1
	10	−12.8	12.3	0.1	21.3	10.6

(A negative value means there was an increase in swelling compared to the baseline.)

a. Compare means and standard deviations by time period for the two treatment groups.

b. Plot these data showing the mean percentage inhibition by time for each treatment group. Does the antihistamine group appear to differ from the placebo group?

18.10 Refer to the data from Exercise 18.9. Give a model for this design, and run a repeated measures analysis of variance to compare the two treatment groups. Do the analysis of variance results agree with your intuition based on the plot of Exercise 18.9?

18.11 Refer to Exercise 18.9. An important question of interest to the researchers is how long after the first dose there is evidence of antihistamine activity. Perform a multiple-comparison procedure to determine the first time at which there is significant evidence of a difference in the mean percentage inhibitions.

Sci.

18.12 There are many running shoes on the market of varying degrees of quality. Long-distance runners require a shoe that provides a significant reduction in impact shock compared to the standard running shoe intended for weekend joggers. A runners' magazine commissioned a study to evaluate three brands of shoes that claim to provide a reduction in impact shock. Ten experienced long-distance runners were selected to participate in the study. The study would consist of placing sensors in the runners' shoes to measure impact forces as the runner ran on a treadmill set at a speed of 4 meters per second. Because the impact force is very dependent on the weight and individual stride of the runner, each of the 10 runners will be observed while using all three brands and a widely sold brand that will serve as a control. The runners were evaluated wearing the four brands in a random order with sufficient time between evaluations to allow the runners to be well rested prior to each evaluation. The impact forces (in Newtons) are presented in the following table with the following notation: BC = control brand and $B1$, $B2$, and $B3$ = three new brands.

Runner	*B*1	*B*2	*B*3	*BC*
1	2,059.3	1,851.6	1,610.9	2,499.9
2	2,663.1	1,442.1	1,145.8	2,075.2
3	2,107.1	1,947.9	1,608.4	2,638.8
4	1,847.7	1,682.5	1,409.8	2,400.2
5	1,875.6	1,743.1	1,419.2	2,389.7
6	1,947.8	1,727.9	1,398.9	2,406.2
7	2,055.8	1,831.9	1,545.5	2,549.3
8	1,747.8	1,571.0	1,185.4	2,307.1
9	1,788.1	1,616.9	1,298.6	2,366.8
10	2,112.9	1,800.0	1,553.6	2,592.3

a. Is there significant ($\alpha = 0.05$) evidence of a difference in the four brands of shoes with respect to their mean peak force?
b. How many runners would be needed to conduct this study as a completely randomized experiment? What would be the gains and losses in conducting the study as a completely randomized design?
c. What conditions are necessary in order for the test conducted in part (a) to provide valid *p*-values?
d. What is the population to which the results of this study can be validly applied?

18.5 Crossover Designs

Psy.

18.13 An investigational drug product was studied under sleep laboratory conditions to determine its effect on duration of sleep. A group of 16 patients willing to participate in the study was randomly assigned to one of two drug sequences; 8 were to receive the investigational drug in period 1 and an identical-appearing placebo in period 2, and the remaining 8 patients were to receive the treatment in the reverse order.

a. Identify the design.
b. Give a model for this design.
c. State the assumptions that might affect the appropriateness of this design.

18.14 Sleep duration data (in hours/night) are shown for the patients of Exercise 18.13.

		Period	
Sequence	Patient	1	2
1	1	8.6	8.0
	2	7.5	7.1
	3	8.3	7.4
	4	8.4	7.3
	5	6.4	6.4
	6	6.9	6.8
	7	6.5	6.1
	8	6.0	5.7
2	9	7.3	7.9
	10	7.5	7.6
	11	6.4	6.3
	12	6.8	7.5
	13	7.1	7.7
	14	8.2	8.6
	15	7.2	7.8
	16	6.7	6.9

Sequence 1 received the investigational drug first and the placebo second; the reverse order applied to sequence 2.

a. Compute means and standard errors per sequence per period.
b. Plot these data to show what happened during the study. Does the investigational drug appear to affect sleep duration? In what way? Use $\alpha = .05$.
c. Run a repeated measures analysis of variance for this design. Draw conclusions. Does the analysis of variance confirm your impressions in part (b)?

18.15 Refer to Exercise 18.13. Suppose we ignore the order in which the patients received the treatments. Count the number of patients who had higher sleep duration on the investigational drug than on placebo.

a. Suggest another simple test for assessing the effectiveness of the investigational drug.
b. Give a p-value for the test of part (a).

18.16 Refer to Exercise 18.13. Suppose the sleep durations for period 2 of sequence 1 were as follows:

8.5 7.6 8.5 8.3 7.2 7.0 6.4 6.1

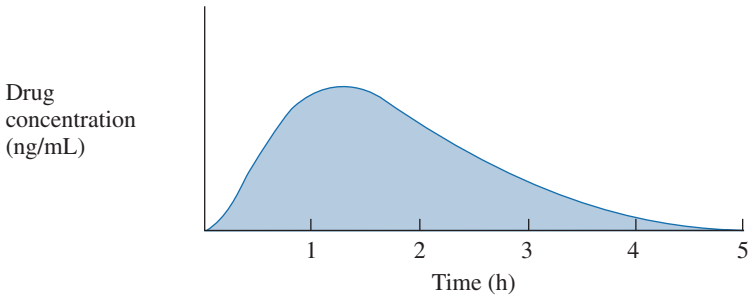
a. Plot the study data for both sequences.
b. Does the design still seem to be appropriate? Is there a possible explanation for what happened?

18.17 Refer to Exercise 18.13. In spite of the results from period 2, we can still get a between-patient comparison of the treatment groups if we use the period 1 results only. Suggest an appropriate test, run the test, and give the p-value for your test. Draw a conclusion.

Med.

18.18 Many of us have been exposed to advertising related to the "bioavailability" of generic and brand-name formulations of the same drug product. One way to compare the bioavailability of two formulations of a drug product is to compare areas under the concentration curve (AUC) for subjects treated with both formulations. For example, the shaded area in the figure represents the AUC for a patient treated with a single dose of a drug.

AUC for a patient treated with a single dose of drug, Exercise 18.18



A three-period crossover design was used to compare the bioavailability of two brand-name (A_1 and A_2) and one generic version (A_3) of weight-reducing agents. Three sequences of administering the drugs were used in the study:

Sequence 1: A_1, A_2, A_3
Sequence 2: A_2, A_3, A_1
Sequence 3: A_3, A_1, A_2

A random sample of five subjects was assigned to each of the three sequences. The AUCs for these 15 patients are shown here.

		Period		
Sequence	**Patient**	**1**	**2**	**3**
1	1	80.2	40.4	38.4
	2	79.1	38.5	36.1
	3	108.4	78.3	56.5
	4	41.2	38.2	26.2
	5	72.7	58.5	36.3

(*continues*)

(*continued*)

Sequence	Patient	Period 1	Period 2	Period 3
2	1	74.6	51.2	48.6
	2	125.3	100.5	86.4
	3	145.5	108.5	96.4
	4	86.7	68.8	58.2
	5	107.8	78.5	53.1
3	1	79.7	40.4	37.2
	2	89.2	68.8	56.2
	3	99.1	76.5	43.9
	4	102.4	88.1	53.4
	5	109.3	98.5	76.8

a. Plot the formulation means (AUCs) by period for each sequence.
b. Is there evidence of a period effect?
c. Do the formulations appear to differ relative to AUC?

18.19 Refer to Exercise 18.18. Run an analysis of variance for a three-period crossover design. Does your analysis confirm the intuition you expressed in Exercise 18.18? Use $\alpha = .05$.

18.20 Refer to Exercise 18.18. Compare the mean AUCs for the three formulations using *only* the period 1 data. Does this analysis confirm the analysis of Exercise 18.19? Why might the analysis of Exercise 18.19 be more suitable or not be more suitable than the "parallel" analysis of this exercise?

Supplementary Exercises

Med.

18.21 The following study is described in ***Chinchilli, Schwab, and Sen (1989).*** The pain of angina is caused by a deficit in oxygen supply to the heart. Calcium channel blockers like verapamil will dilate blood vessels, increasing the supply of blood and oxygen to the heart. This controls chest pain—but only when used regularly. It does not stop chest pain once it starts. The research goal of the study was to assess if there was a difference in four commercial formulations of verapamil (denoted by *A*, *B*, *C*, and *D*). Twenty-six healthy male volunteers were randomly assigned to one of four treatment sequences (*ABCD*, *BCDA*, *CDBA*, or *DABC*). The study protocol required lengthy washouts between treatment periods, and, thus, it was thought that any drug carryover effects from previous time periods would be negligible. The response variable was the area under the plasma time curve (AUC), with values given in the following table.

Subject	Sequence	AUC Period 1	AUC Period 2	AUC Period 3	AUC Period 4	Subject	Sequence	AUC Period 1	AUC Period 2	AUC Period 3	AUC Period 4
1	ABCD	224.29	190.19	135.59	123.19	14	CDAB	399.92	291.57	308.83	301.74
2	BCDA	231.35	265.73	231.22	149.34	15	DABC	117.45	204.20	226.72	127.23
3	CDAB	253.88	202.93	513.31	368.93	16	BCDA	183.20	96.70	200.27	327.96
4	DABC	327.95	453.84	167.11	123.23	17	CDAB	344.18	279.88	317.13	265.73
5	ABCD	326.06	247.43	266.52	212.35	18	DABC	181.75	140.86	254.60	340.48
6	BCDA	259.53	214.41	157.00	188.74	19	ABCD	94.25	58.65	92.93	181.84
7	DABC	347.43	248.74	289.27	329.91	20	BCDA	195.67	297.55	434.38	172.60
8	ABCD	270.10	216.78	273.42	259.00	21	CDAB	458.89	277.73	327.52	345.12
9	BCDA	618.61	401.56	581.72	555.01	22	DABC	383.64	494.78	436.15	380.31
10	CDAB	476.27	210.17	393.30	340.34	23	ABCD	413.53	335.44	291.82	387.86
11	DABC	337.45	169.75	233.68	254.78	24	BCDA	132.88	174.67	105.94	148.22
12	ABCD	483.25	731.50	683.28	366.38	25	CDAB	245.21	142.33	231.53	215.21
13	BCDA	223.04	152.35	107.72	239.81	26	DABC	298.06	324.03	324.13	309.00

a. Plot the formulation means (AUCs) by period for each sequence.
b. Does there appear to be evidence of a period effect?
c. Do the formulations appear to have different AUC means?

18.22 Refer to Exercise 18.21.
a. Write a linear model for the above study. Make sure to identify all parameters in the model.
b. Run an analysis of variance for the data in the study. Does your analysis confirm your intuition expressed in Exercise 18.21?
c. Which pairs of formulations are significantly different?

18.23 Refer to Exercise 18.21. Create a carryover variable as was done in Example 18.7, and conduct a formal test for a significant carryover effect. How are your conclusions altered from the analysis conducted in Exercise 18.22?

18.24 Refer to Exercise 18.21. Using just the period 1 data, test for a difference in the four formulations' mean AUCs. Are your results consistent with the conclusions from Exercise 18.22? Why might the analysis of Exercise 18.22 be more suitable or not be more suitable than the analysis using just the period 1 data?

Med.

18.25 A study was conducted to demonstrate the effectiveness of an investigational drug product in reducing the number of epileptic seizures in patients who have not been helped by standard therapy. Thirty patients participated in the study, with 15 randomized to the drug treatment group and 15 to the placebo group. Patient demographic data are displayed here.

		Group	
		Investigational Drug ($n_1 = 15$)	**Placebo ($n_2 = 15$)**
Age (yr)	Mean (±SD)	37.2 (±10.5)	39.5 (±9.6)
	Range	19–68	21–65
Gender	M	20	16
	F	10	14
Duration of illness (yr)	Mean (±SD)	10.7 (±6.5)	11.5 (±7.3)
	Range	1–18	1–26

a. Do the groups appear to be comparable in terms of these demographic variables?
b. Are the mean ages or durations of illness different? How would you make this comparison?
c. How might you compare the sex distributions of the two groups?

18.26 The seizure data for the study of Exercise 18.25 are shown here. Note that we have baseline seizure rates, as well as seizure rates for 5 months while on therapy.

			Time (months)				
Group	**Patient**	**Baseline**	**1**	**2**	**3**	**4**	**5**
Drug	1	15	11	10	6	5	3
	2	13	6	5	1	2	1
	3	12	8	3	0	3	0

(*continues*)

(*continued*)

			Time (months)				
Group	**Patient**	**Baseline**	**1**	**2**	**3**	**4**	**5**
	4	18	4	2	3	1	2
	5	30	15	14	10	8	20
	6	14	7	9	3	4	1
	7	25	12	18	13	10	6
	8	22	21	18	16	17	25
	9	23	17	14	10	7	1
	10	14	2	1	0	0	0
	11	15	4	5	6	3	2
	12	17	8	7	8	2	6
	13	26	13	10	9	7	4
	14	28	2	1	3	1	3
	15	29	27	29	25	24	22
Placebo	1	16	15	18	14	13	12
	2	18	14	13	12	10	15
	3	14	10	5	4	6	7
	4	19	15	16	9	12	15
	5	12	10	14	16	17	12
	6	11	13	8	7	6	11
	7	31	32	30	21	24	20
	8	32	35	34	31	20	24
	9	21	20	18	15	16	18
	10	26	22	23	21	15	14
	11	13	10	14	12	8	6
	12	17	15	10	3	2	3
	13	18	16	12	14	13	11
	14	23	15	14	18	19	20
	15	10	8	11	10	9	6

a. Plot the mean seizure rates by month for the two groups. Does the investigational drug appear to work?

b. Run a repeated measures AOV, and draw conclusions based on $\alpha = .01$.

18.27 Refer to the data of Exercise 18.26.

a. Consider the change in seizure rates from the baseline to the 5-month reading. Compare the two groups using these data. Do you reach a similar conclusion as was reached in Exercise 18.26?

b. Because seizure rates can be quite variable, some people might compare the maximum change for patients in the two groups. Do these data support your previous conclusions?

Env.

18.28 Gasoline efficiency ratings were obtained on a random sample of 12 automobiles, 6 each of two different models. These ratings were taken at five different times for each of the 12 automobiles.

Model	Car	Time 1	Time 2	Time 3	Time 4	Time 5
1	1	1.43	1.47	1.39	1.40	1.44
1	2	1.50	1.41	1.51	1.53	1.41
1	3	1.79	1.88	1.89	2.00	1.90
1	4	1.87	1.78	2.00	2.00	2.11
1	5	1.85	1.89	1.93	1.86	1.81
1	6	1.89	1.66	1.78	1.77	1.67
2	1	1.63	1.62	1.64	1.63	1.53
2	2	1.81	1.83	1.84	1.83	1.86
2	3	2.25	2.10	2.34	2.27	2.32
2	4	1.79	1.80	1.92	2.03	2.02
2	5	2.11	2.00	2.33	2.46	2.35
2	6	2.10	2.03	2.00	2.09	1.87

a. Compute the mean efficiency for each model at each time point, and plot these data.
b. Draw conclusions from the analysis of variance. Use $\alpha = .05$.
c. What effects, if any, do the Greenhouse-Geisser and Huynh-Feldt correction factors have on the within-model comparisons?

Psy.

18.29 A researcher is designing an experiment in which she plans to compare nine different formulations of a meat product. One factor, F, is percentage of (10%, 15%, and 20%) in the meat. The other factor, C, is cooking method (broil, bake, and fry). She will prepare samples of each of the nine combinations and present them to tasters who will score the samples based on various criteria. Four tasters are available for the study. Each taster will taste nine samples. There are taster-to-taster differences, but the order in which the samples are tasted will not influence the taste scores. The samples will be prepared in the following manner so that the meat samples can be prepared and kept warm for the tasters. A portion of meat containing 10% fat will be divided into three equal portions. Each of the three methods of cooking will then be randomly assigned to one of the three portions. This procedure will be repeated for meat samples having 15% and 20% fat. The nine meat samples will then be tasted and scored by the taster. The whole process is repeated for the other three tasters. The taste scores (0 to 100) are given here.

	10% Fat			15% Fat			20% Fat		
	Broil	Bake	Fry	Broil	Bake	Fry	Broil	Bake	Fry
Taster 1	75	79	82	78	82	81	81	85	87
Taster 2	74	78	81	78	81	83	84	87	88
Taster 3	75	78	79	80	82	83	87	88	92
Taster 4	91	88	83	80	76	73	81	77	74

a. Identify the design.
b. Give an appropriate model with assumptions.
c. Give the sources of variability and degrees of freedom for an AOV.
d. Perform an analysis of variance, and draw conclusions about the effect of fat percentage and method of cooking on the taste of the meat product. Use $\alpha = .05$.

18.30 The following data are from ***Gennings, Chinchilli, and Carter, (1989).*** An in vitro toxicity study of isolated hepatocyte suspensions was conducted to study the impact of combining carbon tetrachloride (CCl_4) and chloroform ($CHCl_3$) on the toxicity of cells. Cell toxicity was measured by the amount of lactic dehydrogenase (LDH) enzyme leakage. The study involved randomly assigning four flasks to each of the 16 treatments obtained by combining four levels of CCl_4 (0, 1.0, 2.5, and 5.0 mM) with four levels of $CHCl_3$ (0, 5, 10, and 25 mM). The percentage of LDH leakage from the cells in each of the

64 flasks was measured just prior to applying the treatment to the flasks and at .01, .25, .5, 1, 2, and 3 hours after applying the treatment. The percentages of LDH leakage are given in the following table.

		Time Since Treatment (hours)									Time Since Treatment (hours)						
CCl_4	$CHCl_3$	0	.01	.25	.5	1.0	2.0	3.0	CCl_4	$CHCl_3$	0	.01	.25	.5	1.0	2.0	3.0
0	0	.08	.09	.09	.08	.10	.10	.12	0	0	.07	.08	.08	.08	.09	.09	.10
0	0	.08	.10	.10	.09	.12	.15	.13	0	0	.06	.08	.06	.07	.08	.10	.11
0	5	.06	.11	.14	.12	.14	.13	.12	0	5	.05	.07	.13	.08	.10	.10	.12
0	5	.11	.14	.16	.18	.20	.21	.14	0	5	.06	.06	.07	.13	.14	.15	.16
0	10	.06	.11	.20	.36	.46	.44	.46	0	10	.06	.07	.17	.18	.21	.22	.22
0	10	.08	.14	.24	.27	.29	.32	.34	0	10	.05	.05	.15	.16	.19	.22	.23
0	25	.07	.10	.25	.51	.65	.66	.70	0	25	.07	.07	.17	.24	.34	.37	.41
0	25	.11	.11	.33	.39	.48	.52	.55	0	25	.07	.06	.16	.24	.31	.36	.41
1	0	.06	.11	.13	.09	.10	.11	.11	1	0	.05	.08	.10	.10	.11	.12	.13
1	0	.08	.14	.15	.14	.16	.19	.21	1	0	.05	.09	.08	.09	.11	.12	.13
1	5	.05	.13	.18	.37	.41	.42	.46	1	5	.06	.10	.14	.16	.16	.20	.18
1	5	.10	.16	.22	.22	.29	.30	.21	1	5	.05	.08	.15	.18	.19	.21	.21
1	10	.06	.10	.25	.61	.57	.60	.63	1	10	.05	.07	.24	.27	.29	.32	.32
1	10	.11	.14	.26	.30	.30	.35	.29	1	10	.05	.06	.16	.21	.24	.27	.27
1	25	.07	.09	.23	.39	.58	.53	.67	1	25	.06	.06	.15	.22	.30	.44	.56
1	25	.08	.11	.28	.40	.42	.75	.72	1	25	.06	.05	.15	.27	.36	.43	.55
2.5	0	.06	.09	.19	.56	.64	.33	.34	2.5	0	.05	.08	.18	.19	.19	.21	.20
2.5	0	.10	.10	.19	.21	.23	.28	.23	2.5	0	.05	.10	.21	.23	.28	.29	.31
2.5	5	.07	.10	.22	.57	.62	.66	.70	2.5	5	.06	.08	.19	.23	.24	.27	.31
2.5	5	.07	.11	.24	.28	.30	.35	.30	2.5	5	.06	.07	.21	.25	.28	.30	.32
2.5	10	.05	.12	.28	.33	.43	.49	.58	2.5	10	.06	.09	.33	.26	.31	.34	.36
2.5	10	.08	.14	.23	.37	.43	.47	.40	2.5	10	.06	.09	.19	.23	.29	.34	.34
2.5	25	.05	.07	.22	.59	.65	.67	.67	2.5	25	.04	.05	.21	.29	.36	.54	.72
2.5	25	.09	.09	.24	.31	.35	.46	.45	2.5	25	.05	.04	.15	.25	.36	.40	.48
5	0	.06	.09	.52	.77	.78	.73	.76	5	0	.06	.08	.45	.50	.49	.60	.71
5	0	.08	.09	.60	.60	.57	.73	.79	5	0	.06	.10	.42	.44	.62	.62	.73
5	5	.05	.11	.21	.27	.30	.36	.41	5	5	.05	.10	.20	.22	.24	.28	.33
5	5	.09	.12	.21	.22	.27	.32	.28	5	5	.05	.08	.17	.21	.26	.27	.32
5	10	.04	.10	.24	.26	.33	.39	.47	5	10	.06	.09	.25	.29	.33	.37	.40
5	10	.11	.11	.23	.27	.31	.36	.31	5	10	.05	.05	.12	.16	.22	.27	.29
5	25	.07	.07	.21	.55	.60	.66	.66	5	25	.05	.05	.23	.31	.35	.53	.66
5	25	.08	.09	.23	.31	.41	.58	.67	5	25	.06	.04	.12	.20	.31	.41	.57

a. Plot the mean percentages of LDH leakage by time for the 16 treatments. Does there appear to be an effect due to increasing the level of CCl_4 or $CHCl_3$?

b. From the plot, does there appear to be an increase in the mean percentages of leakage as time after treatment increases?

c. Plot a profile plot of the mean percentage of LDH leakage separately for each time period. Does there appear to be a difference in the profile plots?

18.31 Refer to Exercise 18.30.

a. Run a repeated measures analysis of variance, and determine if there are significant interaction and/or main effects due to CCl_4 and $CHCl_3$. Is there a significant time effect?

b. Do the conditions necessary for using a split-plot analysis of repeated measures data appear to be valid?

18.32 Refer to Exercise 18.30. Consider as your response variable the proportional change in the mean percentage of leakage at time 3 hours and at time 0. That is,

$$y = \frac{P_3 - P_0}{P_0}$$

where P_0 and P_3 are the percentage of leakage values at times 0 and 3 hours, respectively. Run an analysis of variance on y, and test for significant interaction and/or main effects due to CCl_4 and $CHCl_3$. Do you reach conclusions similar to those obtained in Exercise 18.31?

Engin. **18.33** A group of researchers at a company that produces a leading brand of ice cream design an experiment to evaluate the impact of several artificial sweeteners on the texture of the product. It is well known that replacing natural sweeteners with artificial sweeteners in ice cream can result in a product that is has an unappealing texture. A proposed method to overcome this problem is to increase the blending time in the production process. The researchers decided to use four types of sweeteners: a natural sweetener (control), Aspartame, Saccharin, and Sucralose. Twelve containers of ice cream were made, 3 for each of the four types of sweeteners, with the type of sweetener randomly assigned to the containers. Each of the 12 containers of ice cream was then split into four portions. The four portions were then randomly assigned to one of four blending times: 1 minute, 2 minutes, 5 minutes, and 8 minutes. At the end of the specified blending period, the ice cream was assigned a texture score. The researchers were particularly interested in the impact of the four sweeteners and the blending times on the average texture scores.

		Blending Time(min.)			
Sweetener	**Container**	**1**	**2**	**5**	**8**
Control	1	7	10	17	22
	2	4	4	11	23
	3	4	11	10	31
Aspartame	1	8	12	22	27
	2	6	7	27	30
	3	9	8	29	32
Saccharin	1	7	8	21	35
	2	1	4	13	25
	3	5	4	13	28
Sucralose	1	3	11	21	37
	2	1	12	25	31
	3	4	9	27	32

a. What type of randomization was utilized in this experiment (completely randomized design, randomized complete block design, Latin square design, etc.)?
b. What type of treatment structure was used (single factor, crossed factors, nested factors, etc.)?
c. Identify each of the factors as being fixed or random.
d. Describe the experimental units for each factor and the measurement units.
e. Write a statistical model for this experiment, and include all necessary conditions on the model parameters and variables.

Engin. **18.34** Refer to Exercise 18.33.
a. Do the necessary conditions for testing hypotheses and constructing confidence intervals appear to be satisfied? Justify your answers using the residuals from fitting the model from Exercise 18.33.
b. Construct an ANOVA table for this experiment. Make sure to include expected mean squares and the p-values for the F tests.

c. At the $\alpha = .05$ level, which main effects and interaction effects are significant? Justify your answer by including the relevant p-values.
d. What are your overall conclusions about the impact of the four sweeteners and the blending times on the average texture scores?

Engin. **18.35** Refer to Exercise 18.33.
a. Group the three types of sweeteners along with the control such that all sweeteners in a group are not significantly different from one another with respect to their mean texture scores. Use an experimentwise error rate of $\alpha = .05$.
b. Group the four blending times such that all blending times in a group are not significantly different from one another with respect to their mean texture scores. Use an experimentwise error rate of $\alpha = .05$.
c. In forming the groups in part (a), was it necessary to consider blending times?
d. Provide a 95% confidence interval on the mean texture score for each of four levels of sweetener.
e. Provide a 95% confidence interval on the mean texture score for each of four levels of blending time.

Health **18.36** Sodium nitrate, a preservative that is used in some processed meats, such as bacon, jerky, and luncheon meats, could increase your heart disease risk. A consumer protection organization is evaluating the level of sodium nitrate ($NaNO_2$) from sausages obtained from the three largest food processors—P1, P2, and P3—in the United States. Each manufacturer produces three grades of quality for their sausage—Q1, Q2, and Q3. The processing of different grades of sausage from a common production run may involve different sources of raw materials and processing environments, and these factors sometimes are problematic. Each food processor submits two sausages of each grade from each of three production runs. The amount of $NaNO_2$ is determined and is reported in the following table. The three food processors are the only processors under evaluation, the production runs were randomly selected and are representative of general production runs of each food processor.

	Manufacturer								
	P1 Run			P2 Run			P3 Run		
Grade	R1	R2	R3	R4	R5	R6	R7	R8	R9
Q1	253	265	253	230	234	231	225	228	232
	256	270	251	226	239	232	229	227	232
Q2	262	263	255	257	268	265	277	276	289
	260	266	264	267	258	266	276	277	287
Q3	279	285	277	275	286	284	280	278	282
	279	288	272	272	283	284	276	277	282

a. What type of randomization was utilized in this experiment (completely randomized design, randomized complete block design, Latin square design, etc.)?
b. What type of treatment structure was used (single factor, crossed factors, nested factors, etc.)?
c. Identify each of the factors as being fixed or random.
d. Describe the experimental units for each factor and the measurement units.
e. Write a statistical model for this experiment, and include all necessary conditions on the model parameters and variables.

Health **18.37** Refer to Exercise 18.36.

a. Do the necessary conditions for testing hypotheses and constructing confidence intervals appear to be satisfied?
b. At the $\alpha = .05$ level, which main effects and interactions are significant? Justify your answer by including the relevant p-values.
c. Separate the three quality levels into groups of levels such that all levels in a group are not significantly different from one another with respect to their mean $NaNO_2$ levels. Use an experimentwise error rate of $\alpha = .05$.
d. Provide a 95% confidence interval on the mean $NaNO_2$ level for each of the three quality levels.
e. Provide a 95% confidence interval on the mean $NaNO_2$ level for each of the three food processors.

Health **18.38** Refer to Exercise 18.36.

a. Estimate the size of the variance associated with each of the random factors in the study.
b. Provide the proportions of the total variation associated with each of the sources of random variation in the study.
c. Is the amount of variation in the quantity of $NaNO_2$ across the runs consistent for the three quality levels?

In Exercises 18.39–18.42, describe the experimental situations by provide the following information:

1. Identify the type of randomization (completely randomized design, randomized complete block design, Latin square design, split-plot, crossover, etc.).
2. Identify the type of treatment structure (single factor, crossed factors, nested factors, fractional, etc.).
3. Identify each of the factors as being fixed or random.
4. Describe the experimental units and measurement units.
5. Describe the measurement process: response variable, covariates, subsampling, and repeated measures.
6. Provide a partial AOV table containing just sources of variation and degrees of freedom.

Health **18.39** A research specialist for a large seafood company investigated bacterial growth on oysters and mussels subjected to three different storage temperatures. Nine cold storage units were available. Three storage units were randomly assigned to be used for each of the storage temperatures: 0, 5, and 10°C. Oysters and mussels were stored for 2 weeks in each of the cold storage units. A bacterial count was made from a sample of oysters and a sample of mussels from each storage unit at the end of 2 weeks, so that for each storage unit there is a bacterial value for oysters and a bacterial value for mussels, yielding a total of 18 observations.

Bio. **18.40** A study was designed to compare the effect of a vitamin E supplement on the growth of guinea pigs. There were 15 guinea pigs available for the study. The guinea pigs were randomly assigned to one of the three dose levels of vitamin E with 5 animals per level. For each animal, the body weight was recorded at the end of weeks 1, 3, 4, 5, 6, and 7. All 15 animals were given a growth-inhibiting substance during week 1 and given identical diets during the first four weeks of the study. At the beginning of week 5, the vitamin E treatments were implemented. The three treatment levels (doses of vitamin E) were 0, L (low), and H (high). The data include the response variable WEIGHT for each of the 15 animals for each of the 6 weekly weighings (total of 90 measurements). The other information available for each observation is the levels of DOSE (0, L, and H) and the WEEK (1, 3, 4, 5, 6, and 7). The animals are numbered 1 through 15. In addition, a variable called BEFAFT is created, which has the following values:

BEFAFT = B for weeks 1, 3, and 4—that is, before the start of the vitamin E doses
BEFAFT = A for weeks 5, 6, and 7—that is, after starting the vitamin E doses

Nutrition **18.41** Commercial cheese is manufactured by bacterial fermentation of pasteurized milk. Selected bacteria, referred to as starter cultures, are added to the milk to implement the fermentation. However, some Wild bacteria, nonstarter bacteria, may also be present in cheese,

which may alter the desired quality of the cheese. Thus, cheese manufactured under seemingly identical conditions in two cheese-making facilities may produce cheese of differing quality due to the present of different indigenous nonstarter bacteria. To test the impact of two nonstarter bacteria, R50 and R21, on cheese quality, the nonstarter bacteria were added to the cheese to see if it impacted the quality of the cheese. The researchers decided to use four types of nonstarter bacteria: a control (no nonstarter bacteria added), addition of R50, addition of R21, and addition of a blend of R50 and R21. Twelve containers of cheeses were made, 3 of each of the four types of nonstarter bacteria, with the type of bacteria randomly assigned to the cheese containers. Each of the 12 containers of cheese was then divided into four portions. The four portions were then randomly assigned to one of four aging times: 1 day, 28 days, 56 days, and 84 days. At the end of the specified aging period, the cheese was measured for total free amino acids. The researchers were particularly interested in the bacterial effects and their interaction with aging times.

Engin. **18.42** An industrial engineer is studying the hand insertion of electronic components on printed circuit boards in order to improve the speed of the assembly operation. She has designed three assembly fixtures (F_1, F_2, and F_3) and two workplace layouts (L_1 and L_2) that seem promising. Specialized operators are required to perform the assembly, and it was initially decided to randomly select four operators from the many qualified operators at the plant. However, because the workplaces are in different locations within the plant, it is difficult to use the same operators for each layout. Therefore, the four operators randomly chosen for layout 1 are different individuals from the four operators randomly chosen for layout 2. Each of the operators assembles 4 circuit boards for each of the three fixture types, with the 12 circuit boards assembled in random order. The 96 assembly times are measured in seconds. The engineer is interested in the effects of assembly fixtures (F), workplace layout (L), and operator (O) on the average time required to assemble the circuit boards.

CHAPTER 19

Analysis of Variance for Some Unbalanced Designs

19.1 Introduction and Abstract of Research Study

We examined the analysis of variance for balanced designs in Chapters 8, 14, and 15, where we used appropriate formulas (and corresponding computer solutions) to construct AOV tables and set up hypothesis tests. We also considered another way of performing an analysis of variance. We saw that the sum of squares associated with a source of variability in the analysis of variance table can be found as the drop in the sum of squares for error obtained from fitting reduced and complete models. Although we did not advocate the use of complete and reduced models for obtaining the sums of squares for sources of variability in balanced designs, we did indicate that the procedure was completely general and could be used for any experimental design. In particular, in this chapter, we will make use of complete and reduced models for obtaining the sums of squares in the analysis for *unbalanced designs,* where formulas are no longer readily available and easy to apply.

You might ask why an experimenter would run a study using an unbalanced design, especially since unbalanced designs seem to be more difficult to analyze. In point of fact, most studies do begin by using a balanced design, but for any one of many different reasons, the experimenter is unable to obtain the same number of observations per cell as dictated by the balanced design being employed. Consider a study of three different weight-reducing agents in which five different clinics (blocks)

are employed and patients are to be randomly assigned to the three treatment groups according to a randomized block design. Even if the experimenter plans to have six overweight persons assigned to each treatment at each clinic, the final count will almost certainly show an imbalance of persons assigned to each treatment group. Almost every clinic could be expected to have a few people who would not complete the study. Some people might move from the community, others might drop out due to a lack of efficacy in the program, and so on. In addition, the experimenter might find it impossible to locate 18 overweight people at each clinic who are willing to participate in the study. Because an unbalanced design at the end of a study occurs quite often, we must learn how to analyze data arising from unbalanced designs.

We will next consider a research study in which we are aware of the unbalanced nature of the design prior to running the experiment and hence can design the study to partially accommodate the imbalance so as to minimize any bias with respect to estimating the treatment effects.

Abstract of Research Study: Evaluation of the Consistency of Property Assessors

The county in which a large southwestern city is located received over the past year a large number of complaints concerning the assessed valuation of residential homes. Some of the county residents stated that there was wide variation in residential property valuations depending on which county property assessor determined the property's value. The county employs numerous assessors who determine the value of residential property for the purposes of computing property taxes due from each property owner in the county. The county manager decided to design a study to see whether the assessors differ systematically in their determinations of property values.

The manager needed to determine how to evaluate the consistency in the assessors' determinations of property values. Because the county assessor's office is generally understaffed and the assessors have a complete work schedule, it was decided to randomly select 16 assessors for participation in the study. There is a wide variety in the types of homes and extent of landscaping in the properties throughout the county. This variation in values and styles is thought to be one of the sources of deviations in the assessed valuations of the properties. Thus, the manager carefully selected 16 properties that would represent the wide diversity of properties in the county but all within the midpriced range of homes. To determine consistency, it would be necessary to have the assessors evaluate the same properties, and initially, the study was to have each of the 16 assessors determine a value for each of the 16 properties. This would have required a total of 256 valuations to be done by the 16 assessors. However, this would have been too time consuming. Thus, each assessor was assigned to evaluate 6 of the 16 properties. The necessary number of valuations would be reduced from 256 to 96. The design is a randomized block design with the blocking variable being the 16 properties and the treatment variable being the 16 assessors. Note that the design is no longer a randomized complete block design because each assessor valuated only 6 of the 16 properties. The county statistician was concerned about the incomplete nature of the block design because some of the properties may be more difficult to evaluate than others. Although it would not be possible to have a complete block design, the statistician decided on the following method of assigning the properties to the assessors. We will demonstrate that the design is in fact a **balanced incomplete block design** when we provide the analysis of the research study in Section 19.5.

balanced incomplete block design

Because the design is not a complete block design—only 96 of the 256 possible block–treatment combinations were observed—we cannot use the models and analysis techniques from Chapter 15. The analysis of the research study will be provided in Section 19.5.

19.2 A Randomized Block Design with One or More Missing Observations

unbalanced design

Any time the number of observations is not the same for all factor–level combinations, we call the design **unbalanced.** Thus, a randomized block design or a Latin square design with one or more missing observations is an unbalanced design. We will begin our examination by considering a simple case, a randomized block design with one missing observation.

value of missing observation
estimation bias

The analysis of variance for a randomized block design with one missing observation can be performed rather easily by using the formulas for a randomized complete block design after we have estimated the **value of the missing observation** and corrected for the **estimation bias.**

Let y_{ij} be the response from the experimental unit observed under treatment i in block j. Suppose that the missing observation occurs in cell (k, h), the observation on treatment k in block h. The formula for estimating the missing observation y_{kh} is given by

$$\hat{y}_{kh} = \frac{ty_{k.} + by_{.h} - y_{..}}{(t-1)(b-1)}$$

where t is the number of treatments; b is the number of blocks; $y_{k.}$ is sum of all observations on treatment k, the treatment that has the missing observation; $y_{.h}$ is the sum of all measurements in block h, the block that has the missing observation; and $y_{..}$ is the sum of all the observations.

The sums of squares for the analysis of variance table are obtained by replacing the missing value, y_{kh}, with its estimate, $\hat{y}_{kh}$, and then applying the formulas for a balanced design to the data set that now has no missing cells:

$$\text{TSS} = \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2$$

$$\text{SST} = b\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$\text{SSB} = t\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2$$

$$\text{SSE} = \text{TSS} - \text{SST} - \text{SSB}$$

The value of SST has a bias in its estimation given by

$$\text{Bias} = \frac{(y_{.h} - (t-1)\hat{y}_{kh})^2}{t(t-1)}$$

The corrected treatment sum of squares is $\text{SST}_C = \text{SST} - \text{bias}$. The other sums of squares are given in their uncorrected form.

Another difference in the analysis of variance table for the unbalanced block designs is a change in the entries for degrees of freedom for total and error. Because n in the unbalanced design refers to the number of actual observations, the value of n is

TABLE 19.1
AOV table for testing the effects of treatments with one missing observation

Source	SS	df	MS	F
$\text{Blocks}_{\text{unadj}}$	$\text{SSB}_{\text{unadj}}$	$b - 1$	$\text{MSB}_{\text{unadj}}$	
Treatments_C	SST_C	$t - 1$	MST_C	MST_C/MSE
Error	SSE	$tb - t - b$	MSE	
Total	TSS	$bt - 2$		

given by $n = tb - 1$ due to the missing data point. Therefore, the degrees of freedom for error will be decreased by one to $n - t - b + 1 = tb - t - b$ as compared to $tb - t - b + 1$ for the corresponding balanced design. The AOV table for an unbalanced design with t treatments, b blocks, and one missing value is shown in Table 19.1.

We illustrate the analysis of variance for this design with an example.

EXAMPLE 19.1

Prior to spinning cotton, the cotton must be processed to remove foreign matter and moisture. The most common lint cleaner is the controlled-batt saw-type lint cleaner. Although the controlled-batt saw-type lint cleaner M1 is one of the most highly effective cleaners, it is also one of the cleaners that causes the most damage to the cotton fibers. A cotton researcher designed a study to investigate four alternative methods for cleaning cotton fibers: M2, M3, M4, and M5. Methods M2 and M3 are mechanical, whereas methods M4 and M5 are a combination of mechanical and chemical procedures. The researcher wanted to take into account the impact of different growers on the process and hence obtained bales of cotton from six different cotton farms. The farms will be considered as blocks in the study. After a preliminary cleaning of the cotton, the six bales were thoroughly mixed, and then an equal amount of cotton was processed by each of the five lint-cleaning methods. The losses in weight (in kg) after cleaning the cotton fibers are given in Table 19.2 for the five cleaning methods. During the processing of the cotton samples, the measurements from farm 1 processed by the M1 cleaner were lost.

TABLE 19.2
Measurements of loss (kg) during cotton fiber cleaning

	Farm						
Method	**1**	**2**	**3**	**4**	**5**	**6**	**Mean**
M1	*	6.75	13.05	10.26	8.01	8.42	9.300
M2	5.54	3.53	11.20	7.21	3.24	6.45	6.190
M3	7.67	4.15	9.79	8.27	6.75	5.50	7.022
M4	7.89	1.97	8.97	6.12	4.22	7.84	6.170
M5	9.27	4.39	13.44	9.13	9.20	7.13	8.760
Mean	7.593	4.158	11.290	8.198	6.280	7.068	7.426

Estimate the value for the missing observation and then perform an analysis of variance to test for differences in the mean weight losses for the five methods of cleaning cotton fibers.

Solution For this randomized block design, $b = 6$ and $t = 5$ with one missing value in cell $(1, 1)$. Therefore, we need to compute the following values:

$$y_{1.} = \text{sum of all measurements on method M1}$$
$$= 6.75 + 13.05 + 10.26 + 8.01 + 8.42 = 46.49$$

$$y_{.1} = \text{sum of all measurements on batch 1}$$
$$= 5.54 + 7.67 + 7.89 + 9.27 = 30.37$$
$$y_{..} = \text{sum of all measurements}$$
$$= 6.75 + 13.05 + \cdots + 7.13 = 215.36$$

The estimate of the missing value, y_{11}, is given by

$$\hat{y}_{11} = \frac{ty_{1.} + by_{.1} - y_{..}}{(t-1)(b-1)} = \frac{5(46.49) + 6(30.37) - 215.36}{(5-1)(6-1)}$$
$$= \frac{199.31}{20} = 9.9655$$

Replacing the missing value with its estimate, 9.9655, we next compute the sum of squares using the formulas of Chapter 15 for a balanced randomized block design with $t = 5$ and $b = 6$. First, we obtain the treatment and farm means (with the missing value replaced with 9.9655), as shown in Table 19.3.

TABLE 19.3
Method and batch means

Method Mean	Farm Mean
$\bar{y}_{1.} = 9.409$	$\bar{y}_{.1} = 8.067$
$\bar{y}_{2.} = 6.190$	$\bar{y}_{.2} = 4.158$
$\bar{y}_{3.} = 7.022$	$\bar{y}_{.3} = 11.290$
$\bar{y}_{4.} = 6.170$	$\bar{y}_{.4} = 8.198$
$\bar{y}_{5.} = 8.760$	$\bar{y}_{.5} = 6.280$
	$\bar{y}_{.6} = 7.068$
Overall mean	$\bar{y}_{..} = 7.511$

Note that the means for method 1 and farm 1 and the overall mean incorporate the estimated value for the missing observation. We next obtain the four sums of squares.

$$\text{TSS} = \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2$$
$$= (9.9655 - 7.511)^2 + (6.75 - 7.511)^2 + \cdots + (7.13 - 7.511)^2 = 219.887$$
$$\text{SST} = b\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2$$
$$= 6\,[(9.409 - 7.511)^2 + (6.190 - 7.511)^2 + (7.022 - 7.511)^2$$
$$+ (6.170 - 7.511)^2 + (8.760 - 7.511)^2] = 53.624$$
$$\text{SSB} = t\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2$$
$$= 5\,[(8.067 - 7.511)^2 + (4.158 - 7.511)^2 + (11.290 - 7.511)^2$$
$$+ (8.198 - 7.511)^2 + (6.280 - 7.511)^2 + (7.068 - 7.511)^2] = 140.032$$
$$\text{SSE} = \text{TSS} - \text{SST} - \text{SSB} = 219.887 - 53.624 - 140.032 = 26.231$$
$$\text{Bias} = \frac{(y_{.1} - (t-1)\hat{y}_{11})^2}{t(t-1)} = \frac{[30.37 - (5-1)9.9655]^2}{5(5-1)} = 4.5049$$
$$\text{Corrected treatment SS} = \text{SST}_C = \text{SST} - \text{bias} = 53.624 - 4.5049 = 49.119$$

The AOV table for Example 19.1 is shown in Table 19.4.

TABLE 19.4
AOV table for testing the effects of treatments with one missing observation

Source	SS	df	MS	F	p-value
$\text{Blocks}_{\text{unadj}}$	140.032	5	28.01		
$\text{Treatments}_{\text{C}}$	49.119	4	12.28	7.96	.0008
Error	26.231	19	1.543		
Total	219.887	28			

The F test for a significant difference in the five method means is highly significant (p-value = .0008). The mean losses in cotton fiber were somewhat higher when using methods 1 and 5 in comparison to the other three methods. ■

comparisons among treatment means

Having seen an analysis of variance, we may wish to make certain **comparisons among the treatment means.** We'll run pairwise comparisons using the Tukey-Kramer W procedure. The value of W for comparing the treatment with a missing observation and any other treatment mean is

$$W^* = \frac{q_\alpha(t, v)}{\sqrt{2}}\sqrt{\text{MSE}\left(\frac{2}{b} + \frac{t}{b(b-1)(t-1)}\right)}$$

For any pair of treatments with no missing value, the least significant difference is as before—namely,

$$W = q_\alpha(t, v)\sqrt{\frac{\text{MSE}}{b}}$$

EXAMPLE 19.2

In Example 19.1, we found that there was significant evidence of a difference in the mean loss in cotton fiber for the five methods. The researchers would like to determine which pairs of methods have differences. Run a pairwise comparison of the five methods using the Tukey-Kramer W procedure.

Solution Example 19.1 involved a study in which the design was a randomized block design with $t = 5$ treatments and $b = 6$ blocks. There was a single missing observation. From Table 19.4, we have MSE = 1.543 with 19 degrees of freedom. Using $\alpha = .05$, the value of W for comparing the method with the missing observation, method 1, with the other four methods is computed as

$$W^* = \frac{q_{.05}(5, 19)}{\sqrt{2}}\sqrt{\text{MSE}\left(\frac{2}{b} + \frac{t}{b(b-1)(t-1)}\right)}$$

$$= 3.01\sqrt{1.543\left(\frac{2}{6} + \frac{5}{6(6-1)(5-1)}\right)} = 2.289$$

For comparing any pair not including method 1, the value of LSD is

$$W = q_{.05}(5, 19)\sqrt{\frac{\text{MSE}}{b}} = 4.25\sqrt{\frac{1.543}{6}} = 2.155$$

Using the two values of W, we obtain the results shown in Table 19.5, with the mean for method 1 computed using the estimated missing observation.

TABLE 19.5
Paired comparison of five methods

Pair Compared	Difference in Means	W	Conclusion
M1 & M2	9.409 − 6.190 = 3.219	2.289	Significant
M1 & M3	9.409 − 7.022 = 2.387	2.289	Significant
M1 & M4	9.409 − 6.170 = 3.239	2.289	Significant
M1 & M5	9.409 − 8.760 = .649	2.289	Not Significant
M2 & M3	6.190 − 7.022 = −.832	2.155	Not Significant
M2 & M4	6.190 − 6.170 = .020	2.155	Not Significant
M2 & M5	6.190 − 8.760 = −2.570	2.155	Significant
M3 & M4	7.022 − 6.170 = .852	2.155	Not Significant
M3 & M5	7.022 − 8.760 = −1.738	2.155	Not Significant
M4 & M5	6.170 − 8.760 = −2.590	2.155	Significant

We can group the five methods on the basis of an LSD pairwise comparison as follows:

Method 1	Method 5	Method 3	Method 2	Method 4
9.409	8.760	7.022	6.190	6.170
a	*ab*	*bc*	*c*	*c*

fitting complete and reduced models

The formulas for estimating missing observations in a randomized block design become more complicated with more missing data, as do the formulas for *W*s. Because of this, we will consider **fitting complete and reduced models** to analyze unbalanced designs. We will illustrate the procedure first by examining an unbalanced randomized block design.

Because it would require more data input for a computer solution using the general linear model format with dummy variables presented in Chapter 12, we will represent the complete and reduced models for testing treatments as follows:

$$\text{Complete model (model 1):} \quad y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

$$\text{Reduced (model 2):} \quad y_{ij} = \mu + \beta_j + \varepsilon_{ij}$$

where β_j is the jth block effect and τ_i is the ith treatment effect.

By fitting model 1 (using SAS or other computer software), we obtain SSE_1. Similarly, a fit of model 2 yields SSE_2. The difference in the two sums of squares for error, $SSE_2 - SSE_1$, gives the drop in the sum of squares due to treatments. Because this is an unbalanced design, the block effects do not cancel out when comparing treatment means as they do in a balanced randomized block design (see Chapter 15). The difference in the sums of squares, $SSE_2 - SSE_1$, has been adjusted for any effects due to blocks caused by the imbalance in the design. This difference is called the **sum of squares due to treatments adjusted for blocks.**

SST_{adj}

$$SSE_2 - SSE_1 = SST_{adj}$$

The sum of squares due to blocks **unadjusted for any treatment differences** is obtained by subtraction:

$$SSB = TSS - SST_{adj} - SSE$$

where SSE and TSS are sums of squares from the complete model. (*Note:* We could also obtain SSB, the uncorrected sum of squares for blocks, using the formula of Section 15.2).

AOV table, treatments

The **analysis of variance table for testing the effect of treatments** is shown in Table 19.6. In the table, n is the number of actual observations.

TABLE 19.6
AOV table for testing the effects of treatments, unbalanced randomized block design

Source	SS	df	MS	F
Blocks	SSB	$b - 1$	—	—
Treatments_{adj}	SST_{adj}	$t - 1$	MST_{adj}	MST_{adj}/MSE
Error	SSE	$n - b - t + 1$	MSE	
Total	TSS	$n - 1$		

TABLE 19.7
AOV table for testing effects of blocks, unbalanced randomized block design

Source	SS	df	MS	*F*
$\text{Blocks}_{\text{adj}}$	SSB_{adj}	$b - 1$	MSB_{adj}	$\text{MSB}_{\text{adj}}/\text{MSE}$
Treatments	SST	$t - 1$	—	—
Error	SSE	$n - t - b + 1$	MSE	—
Total	TSS	$n - 1$		

The corresponding sum of squares for testing the effect of blocks has the same complete model (model 1) as before, and

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

SSB_{adj}

is the reduced model (model 2). The sum of squares drop, $\text{SSE}_2 - \text{SSE}_1 = \text{SSB}_{\text{adj}}$, is the **sum of squares due to blocks after adjusting for the effects of treatments.** By subtraction, we obtain

$$\text{SST} = \text{TSS} - \text{SSB}_{\text{adj}} - \text{SSE}$$

The AOV table is shown in Table 19.7.

Note that SST and SST_{adj} are not the same quantity in an unbalanced design; they will be the same only for a balanced design. Similarly, SSB and SSB_{adj} are different quantities in an unbalanced design. For an unbalanced design, we have the following identities:

$$\text{TSS} = \text{SST}_{\text{adj}} + \text{SSB} + \text{SSE} = \text{SST} + \text{SSB}_{\text{adj}} + \text{SSE}$$

but

$$\text{TSS} \neq \text{SST}_{\text{adj}} + \text{SSB}_{\text{adj}} + \text{SSE}$$

EXAMPLE 19.3

Use the data in Example 19.1 to obtain the sum of squares due to treatments after adjusting for the effects of blocks and the sum of squares due to blocks after adjusting for the effects of treatments by using the full versus reduced models technique. Compare your answers to the calculations from Example 19.1.

Solution The following output from Minitab was obtained from fitting the following three models:

Model 1, complete model: $y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$
Model 2, reduced model for treatments: $y_{ij} = \mu + \beta_j + \varepsilon_{ij}$
Model 3, reduced model for blocks: $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

```
Model 1: Analysis of Variance for Loss, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS  Adj MS      F      P
Batch    5  138.304  139.661  27.932  20.23  0.000
Method   4   49.120   49.120  12.280   8.90  0.000
Error   19   26.230   26.230   1.381
Total   28  213.653

Model 2: Analysis of Variance for Loss, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS  Adj MS      F      P
Batch    5  138.304  138.304  27.661   8.44  0.000
Error   23   75.349   75.349   3.276
Total   28  213.653

Model 3: Analysis of Variance for Loss, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS  Adj MS      F      P
Method   4   47.763   47.763  11.941   1.73  0.177
Error   24  165.891  165.891   6.912
Total   28  213.653
```

To obtain the sum of squares due to methods after adjusting for the effects of farm, we use the sum of squares for error from models 1 and 2:

$$\text{SST}_{\text{adj}} = \text{SSE}_2 - \text{SSE}_1 = 75.349 - 26.230 = 49.119$$

This is the same value that we obtained in Example 19.1 by using the formulas. The F test for comparing the five method means is given by

$$F = \frac{\text{SST}_{\text{adj}}/(t-1)}{\text{MSE}_1} = \frac{49.119/4}{1.381} = 8.89 \text{ with } p\text{-value} = Pr[F_{4,19} \geq 8.89] = .0003$$

To obtain the sum of squares due to farm after adjusting for the effects of methods, we use the sum of squares for error from models 1 and 3:

$$\text{SSB}_{\text{adj}} = \text{SSE}_3 - \text{SSE}_1 = 165.891 - 26.230 = 139.66$$

The F test for comparing the six farm means is given by

$$F = \frac{\text{SSB}_{\text{adj}}/(b-1)}{\text{MSE}_1} = \frac{139.66/5}{1.381} = 20.23 \text{ with } p\text{-value} = Pr[F_{5,19} \geq 20.23] < .0001$$

Thus, there is a very significant difference in the farm means and in the method means. ■

19.3 A Latin Square Design with Missing Data

Recall that a $t \times t$ Latin square design can be used to compare t treatment means while filtering out two additional sources of variability (rows and columns). The treatments are randomly assigned in such a way that each treatment appears in every row and in every column. In this section, we will illustrate the method for performing an analysis of variance in a Latin square design when one observation is missing. Then we will use the general method of fitting complete and reduced models with missing observations, described for the randomized block design in Section 19.2, for more complicated designs.

Let y_{ijk} be the response from the experimental unit observed in the ith row and jth column receiving treatment k. Suppose that the missing observation occurs in cell (g, h, m), the response from the experimental unit observed in the gth row and hth column receiving treatment m. The formula for **estimating a single missing observation,** y_{ghm}, in a $t \times t$ Latin square is given by

estimating missing value

$$\hat{y}_{ghm} = \frac{t(y_{g..} + y_{.h.} - y_{..m}) + 2y_{...}}{(t-1)(t-2)}$$

where $y_{g..}$ is the sum of all observations in the gth row, $y_{.h.}$ is the sum of all observations in the hth column, $y_{..m}$ is the sum of all observations receiving the mth treatment, $y_{...}$ is the sum of all $n = t^2 - 1$ observations, and t is the number of treatments in the Latin square.

The sums of squares for the analysis of variance table are obtained by replacing the missing value, y_{ghm}, with its estimate, $\hat{y}_{ghm}$, and then applying the formulas for a balanced design to the data set that now has no missing cells:

$$\text{TSS} = \sum_{i=1}^{t}\sum_{j=1}^{t}(y_{ijk} - \bar{y}_{...})^2$$

$$\text{SST} = t\sum_{k=1}^{t}(\bar{y}_{..k} - \bar{y}_{...})^2$$

TABLE 19.8
AOV table for a Latin square design with one missing value

Source	SS	df	MS	*F*
Row	SSR	$t - 1$	MSR	—
Column	SSC	$t - 1$	MSC	—
Treatment	SST_C	$t - 1$	MST_C	MST_C/MSE
Error	SSE	$n - 3t + 2$	MSE	
Total	TSS	$n - 1$		

$$SSR = t\sum_{i=1}^{t}(\bar{y}_{i..} - \bar{y}_{...})^2$$

$$SSC = t\sum_{j=1}^{t}(\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$SSE = TSS - SST - SSR - SSC$$

The mean square for treatment is a biased estimator for the expected mean square treatment in a balanced Latin square, $\sigma_\varepsilon^2 + t\theta_\tau$. An estimator of this bias is given by

$$\text{Bias} = \left(\frac{y_{...} - y_{g..} - y_{.h.} - (t-1)y_{..m}}{(t-1)(t-2)}\right)^2$$

The corrected treatment sum of squares is

$$SST_C = SST - \text{bias}$$

The other sums of squares are given in their uncorrected form. This results in $MST_C = SST_C/(t-1)$ being an unbiased estimator of $\sigma_\varepsilon^2 + t\theta_\tau$. With $n = t^2 - 1$, the number of observed data values in the Latin square design, we obtain the AOV table shown in Table 19.8 for the Latin square design with the one missing observation estimated by $\hat{y}_{ghm}$.

EXAMPLE 19.4

A company has considered the properties (such as strength, elongation, and so on) of many different variations of nylon stockings in trying to select the experimental stockings to be the subject of extensive consumer acceptance surveys.

Five versions (A, B, C, D, and E) of the stockings have passed the preliminary screening and are scheduled for more extensive testing. As part of the testing, five samples of each type are to be examined for elongation under constant stress by each of five investigators on five separate days. The analyses are to be performed following the random assignment of a Latin square. The elongation data (in centimeters) are displayed in Table 19.9.

TABLE 19.9
Elongation data for Example 19.4

	Day									
Investigator	1		2		3		4		5	
1	B	22.1	A	18.6	C	23.0	E	24.3	D	17.1
2	C	23.5	D	16.5	A	18.7	B	22.0	E	*M*
3	D	17.4	E	23.8	B	22.8	C	23.9	A	20.0
4	A	20.3	B	23.4	E	25.9	D	18.7	C	24.2
5	E	25.7	C	24.8	D	18.9	A	20.6	B	24.6

Note that the measurement on variety E stockings for investigator 2 is missing and that the experiment was not rerun to obtain an observation. Use the methods of this section to estimate the missing value.

Solution For our data, the treatment, row, and column totals corresponding to the missing observation, and the overall total are

$$y_{..5} = 99.70 \qquad y_{2..} = 80.70 \qquad y_{.5.} = 85.90 \qquad y_{...} = 520.80$$

Then with $t = r = c = 5$, we find

$$\hat{y}_{255} = \frac{5(80.70 + 85.90 + 99.70) - 2(520.80)}{(5-1)(5-2)} = 24.1583$$

We will replace the missing observation with its least-squares estimate, $\hat{y}_{255}$, and compute the sums of squares using the formulas for a complete 5×5 Latin square. The investigator, day, and version sample means are shown in Table 19.10.

TABLE 19.10
Sample means for Example 19.4

Investigator	Day	Version	Overall
$\bar{y}_{1..} = 21.020$	$\bar{y}_{.1.} = 21.800$	$\bar{y}_{..1} = 19.640$	$\bar{y}_{...} = 21.79833$
$\bar{y}_{2..} = 20.97166$	$\bar{y}_{.2.} = 21.420$	$\bar{y}_{..2} = 22.980$	
$\bar{y}_{3..} = 21.580$	$\bar{y}_{.3.} = 21.860$	$\bar{y}_{..3} = 23.880$	
$\bar{y}_{4..} = 22.500$	$\bar{y}_{.4.} = 21.900$	$\bar{y}_{..4} = 17.720$	
$\bar{y}_{5..} = 22.920$	$\bar{y}_{.5.} = 22.01166$	$\bar{y}_{..5} = 24.77166$	

$$\begin{aligned}
\text{TSS} &= (22.1 - 21.79833)^2 + (18.6 - 21.79833)^2 + \cdots + (24.6 - 21.79833)^2 \\
&= 197.20 \\
\text{SSR} &= 5\{(21.020 - 21.79833)^2 + (20.97166 - 21.79833)^2 + \cdots + (22.920 \\
&\quad - 21.79833)^2\} = 15.44 \\
\text{SSC} &= 5\{(21.8 - 21.79833)^2 + (21.42 - 21.79833)^2 + \cdots + (22.01166 \\
&\quad - 21.79833)^2\} = 1.01 \\
\text{SST} &= 5\{(19.64 - 21.79833)^2 + (22.98 - 21.79833)^2 + \cdots + (24.77166 \\
&\quad - 21.79833)^2\} = 179.31 \\
\text{SSE} &= 197.20 - 15.44 - 1.01 - 179.31 = 1.44 \\
\text{Bias} &= \left(\frac{520.80 - 80.70 - 85.90 - (5-1)99.70}{(5-1)(5-2)}\right)^2 = 13.82 \\
\text{Corrected treatment} &= \text{SST}_C = 179.31 - 13.82 = 165.49
\end{aligned}$$

The analysis of variance table for this study is given in Table 19.11.

TABLE 19.11
AOV table for Example 19.4

Source	SS	df	MS	*F*
Investigator	15.44	4	3.86	—
Day	1.01	4	.25	—
Version	165.49	4	41.37	316.04
Error	1.44	11	.13	
Total	197.20	23		

■

Having located a significant effect due to treatments, we can make pairwise treatment comparisons using the following formulas. The Tukey-Kramer W for comparing the treatment with the missing value and any other treatment is

$$W^* = \frac{q_\alpha(t, v)}{\sqrt{2}}\sqrt{\text{MSE}\left(\frac{2}{t} + \frac{1}{(t-1)(t-2)}\right)}$$

For any other pair of treatments, the LSD is as before:

$$W = q_\alpha(t, v)\sqrt{\frac{\text{MSE}}{t}}$$

The value for MSE is taken from the analysis of variance table.

EXAMPLE 19.5

Refer to Example 19.4.

a. Test for a significant difference in the mean elongations of the five versions of the stockings.
b. Determine which pairs of the five versions of the stockings are significantly different.

Solution

a. We want to test the hypotheses H_0: $\mu_A = \mu_B = \mu_C = \mu_D = \mu_E$ versus H_a: Not all μs are equal. The test statistic for testing for differences in the mean elongations is given by

$$F = \frac{\text{SST}_C/(t-1)}{\text{SSE}/(n-3t+2)} = \frac{165.49/4}{1.44/11} = 316.04$$

using the values from Table 19.11. The F test has p-value $= Pr(F_{4,11} \geq 316.04) < .0001$. Therefore, we conclude that there is significant evidence of a difference in mean elongations of the five versions of the stockings.

b. For comparing pairs of versions of the stockings that do not having missing observations, we will use

$$W = q_{.05}(5, 11)\sqrt{\frac{\text{MSE}}{t}} = 4.57\sqrt{\frac{.131}{5}} = .740$$

For comparing pairs of versions of the stockings that have missing observations, we will use

$$W^* = \frac{q_{.05}(5, 11)}{\sqrt{2}}\sqrt{\text{MSE}\left(\frac{2}{t} + \frac{1}{(t-1)(t-2)}\right)}$$

$$= 3.231\sqrt{(.131)\left(\frac{2}{5} + \frac{1}{(5-1)(5-2)}\right)} = .813$$

Using the two values of LSD, we obtain the results shown in Table 19.12, with the mean for method 1 computed using the estimated missing observation.

TABLE 19.12
Paired comparison of five versions

Pair Compared	Difference in Means	W	Conclusion
A & B	$19.64 - 22.98 = -3.34$	.740	Significant
A & C	$19.64 - 23.88 = -4.24$	.740	Significant
A & D	$19.64 - 17.72 = 1.92$	.740	Significant
A & E	$19.64 - 24.77 = -5.13$	.813	Significant

(*continues*)

TABLE 19.12 (*continued*)

Pair Compared	Difference in Means	W	Conclusion
B & C	22.98 − 23.88 = −.90	.740	Significant
B & D	22.98 − 17.72 = 5.26	.740	Significant
B & E	22.98 − 24.77 = −1.79	.813	Significant
C & D	23.88 − 17.72 = 6.16	.740	Significant
C & E	23.88 − 24.77 = −.89	.813	Significant
D & E	17.72 − 24.77 = −7.05	.813	Significant

The treatment sample means and comparisons are given in Table 19.13.

TABLE 19.13 Results of paired comparisons

Version	D	A	B	C	E*
Mean	17.72	19.64	22.98	23.88	24.77
Grouping	*a*	*b*	*c*	*d*	*e*

*Version E is missing an observation

All pairs of versions of the stockings have significantly different mean elongations. ■

fitting full and reduced models

For Latin square designs with more than one missing observation, it is easier to use the method of **fitting full and reduced models** to adjust the treatment sum of squares for imbalances in the design due to missing observations. The complete model is given by

$$\text{Model 1:} \quad y_{ijk} = \mu + \tau_k + \beta_i + \gamma_j + \varepsilon_{ijk}$$

where y_{ijk} is the observation in the ith row and jth column on treatment k. This model is fit to the observed data without estimating the missing values. We obtain the error sum of squares, which we will denote as SSE_1. Next, we fit the reduced model without the treatment effect,

$$\text{Model 2:} \quad y_{ijk} = \mu + \beta_i + \gamma_i + \varepsilon_{ijk}$$

to the observed data without estimating the missing values. We again obtain an error sum of squares, which we will denote as SSE_2. The difference in these two error sums of squares is the corrected sum of squares for treatments:

$$\text{SST}_C = \text{SSE}_2 - \text{SSE}_1$$

The test for treatment effects is the F test given in Table 19.8:

$$F = \frac{\text{SST}_C/(t-1)}{\text{SSE}_1/(n-3t+2)}$$

where n is the number of observed data values. We could obtain the corrected sums of squares for row and column effects in a similar fashion. By fitting a reduced model including the treatment effect and row effect but without the column effect, we could obtain the sum of squares error for needed to obtain the adjusted column effect. Similarly, we could obtain the adjusted row effect. In most cases, the test for significant column or row effects is not of interest.

EXAMPLE 19.6

Refer to Example 19.4.

Use the following output to compute the sums of squares for version of stockings and error. Compare these values to the values computed using the estimated missing value formulas. The output was obtained without replacing the missing value with its estimate.

```
The SAS System  GLM Procedure

Class        Levels     Values
INVEST            5     1 2 3 4 5
DAY               5     1 2 3 4 5
VERSION           5     A B C D E

Number of Observations Read          25
Number of Observations Used          24

Model 1: Full Model

Dependent Variable: ELONG

                                  Sum of
Source                DF         Squares   Mean Square   F Value   Pr > F
Model                 12     189.9568333    15.8297361    120.66   <.0001
Error                 11       1.4431667     0.1311970
Corrected Total       23     191.4000000

Source                DF     Type III SS   Mean Square   F Value   Pr > F
INVEST                 4      14.3688333     3.5922083     27.38   <.0001
DAY                    4       0.9428333     0.2357083      1.80   0.1998
VERSION                4     165.4943333    41.3735833    315.35   <.0001

Model 2: Reduced Model

Dependent Variable: ELONG

                                  Sum of
Source                DF         Squares   Mean Square   F Value   Pr > F
Model                  8      24.4625000     3.0578125      0.27   0.9646
Error                 15     166.9375000    11.1291667
Corrected Total       23     191.4000000

Source                DF     Type III SS   Mean Square   F Value   Pr > F
INVEST                 4     23.49000000    5.87250000      0.53   0.7172
DAY                    4      2.13400000    0.53350000      0.05   0.9952
```

SSE = 1.44 with df = 11. $SST_{adj} = SSE_{reduced} - SSE_{complete} = 166.9375 - 1.4432 = 165.4943$ with df = 15 − 11 = 4.

These are the same values that we obtained in Example 19.4 using the estimated missing value formulas. ■

19.4 Balanced Incomplete Block (BIB) Designs

The designs we have discussed thus far in this chapter were unbalanced due to unforeseen circumstances caused by some accident while conducting the experiment or during data processing. Sometimes, however, we may be forced to design an experiment in which we must sacrifice some balance in order to perform the experiment. This often occurs when the number of experimental units per block is fewer than the number of treatments under consideration. Consider the following example.

EXAMPLE 19.7

Suppose the quality control laboratory of a chemical company needs to evaluate five different formulations (A, B, C, D, and E) of a paint for consistency of color. Four samples of each formulation are evaluated on a daily basis. The laboratory has five technicians available for running the tests, and each technician can evaluate at most four samples per day. Thus, it is not possible to conduct a randomized complete block design because every formulation cannot be evaluated by every technician.

However, it may be possible to achieve a partial balance in the design by having each pair of formulations evaluated by the same number of technicians. Display the treatment assignments to achieve this partial balance in the design.

Solution The treatment assignments are displayed in Table 19.14.

TABLE 19.14 Assignment of formulations to quality control technicians

Technician	Formulation			
1	D	B	A	E
2	E	A	C	D
3	A	C	D	B
4	C	E	B	A
5	B	D	E	C

Note that each pair of formulations is evaluated by three technicians. ■

Any randomized block design in which the number of treatments t to be investigated is larger than the number of experimental units available per block is called an **incomplete block design.** Thus, whenever homogeneous blocks of $k < t$ experimental units exist or can be constructed, an incomplete block design cannot be avoided. However, it may be possible to achieve partial balance in the design. One such incomplete block design is defined here.

DEFINITION 19.1

A **balanced incomplete block (BIB) design** is an experimental design in which there are t treatments assigned to b blocks such that

1. Each block contains $k < t$ experimental units.
2. Each treatment appears at most once in each block.
3. Each block contains k treatments.
4. Every treatment appears in exactly r blocks.
5. Every pair of treatments occurs together in λ blocks.

From Definition 19.1, we can conclude that for a design to be a BIB design,

- Every pair of treatments appears in the same block equally often.
- Each treatment is observed r times.
- The number of observations, n, must satisfy $n = rt = kb$.
- $\lambda < r < b$.
- $\lambda = r(k-1)/(t-1)$ must be an integer.

EXAMPLE 19.8

Refer to Example 19.7. Verify that the design displayed in Table 19.14 satisfies the conditions for a BIB design.

Solution We had $b = 5$ blocks (technicians) and $t = 5$ treatments (formulations). There were $k = 4$ treatments per block; hence, $k = 4 < 5 = t$, which results in an incomplete block design. Now, each formulation appeared in exactly $r = 4$ blocks. For the design to be a BIB design, we would need to have every pair of formulations evaluated by $\lambda = r(k-1)/(t-1) = 4(4-1)/(5-1) = 3$ technicians. Examining the assignment of technicians to formulations in Table 19.14, we find that each pair of formulations is evaluated by three technicians. Thus, the design given in Table 19.14 is a BIB design. ■

In many situations, we do not have complete flexibility in designing an experiment because a BIB design does not exist for all possible choices of t, k, b, and r. For example, suppose we have $t = 6$ treatments to be investigated and $b = 4$ blocks, each containing $k = 3$ experimental units. Thus, each treatment could be observed $r = 2$ times. However, for the design to be a BIB design, $\lambda = r(k-1)/(t-1)$ would have to be an integer. In fact, however, $\lambda = 2(3-1)/(6-1) = 4/5$, which is obviously not an integer. Thus, a BIB design cannot be constructed for this combination of treatments and blocks. There are procedures for constructing BIB designs and more-complicated incomplete block designs. The books by Cochran and Cox (1957), Lentner and Bishop (1993), and Kuehl (2000) contain tables of BIB designs and methods for constructing such designs. Several statistical software programs (SAS and Minitab, for example) will construct BIB designs for specified values of t, k, b, and r.

The analysis of variance for a balanced incomplete block design can be performed either by using specifically developed formulas or by using the method of fitting complete and reduced models as discussed for unbalanced designs. We will present the shortcut formulas for the analysis of variance table shown in Table 19.15.

The model for a BIB design is given here:

$$y_{ijg} = \mu + \tau_i + \beta_j + \varepsilon_{ijg} \qquad \text{for } i = 1, \ldots, t; \quad j = 1, \ldots, b; \quad g = \delta_{ij}$$

where δ_{ij} is 1 if the ith treatment appears in the jth block and is 0 otherwise. The terms in the model are μ, the overall mean; τ_i, the ith treatment effect; and β_j, the jth block effect. The ε_{ijg}s are independent and normally distributed with mean 0 and variance σ_ε^2. From this model, we compute the sum of squares for blocks, unadjusted for treatments (SSB) and the total sum of squares (TSS) as we did previously:

$$\text{TSS} = \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2$$

where $n = rt = bk$ is the actual number of data values and

$$\text{SSB} = k\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2$$

where $\bar{y}_{.j}$ is the mean of all observations in the jth block and $\bar{y}_{..}$ is the overall mean. Then, if we define

$y_{i.}$ = sum of all observations on treatment i

$B_{(i)}$ = sum of all measurements for blocks that contain treatment i

the sum of squares for treatments adjusted for blocks is

$$\text{SST}_{\text{adj}} = \frac{t-1}{nk(k-1)}\sum_{i}(ky_{i.} - B_{(i)})^2$$

TABLE 19.15
Analysis of variance table for a balanced incomplete block design

Source	SS	df	MS	F
Blocks	SSB	$b-1$	—	—
$\text{Treatments}_{\text{adj}}$	SST_{adj}	$t-1$	MST_{adj}	$\text{MST}_{\text{adj}}/\text{MSE}$
Error	SSE	$n-t-b+1$	MSE	
Total	TSS	$n-1$		

The sum of squares for error is found by subtraction:

$$SSE = TSS - SSB - SST_{adj}$$

As indicated in Table 19.15, the test statistic for testing the hypothesis of no difference among the treatment means is MST_{adj}/MSE.

EXAMPLE 19.9

A large company enlisted the help of a random sample of 12 potential consumers in a given geographical location to compare the physical characteristics (such as firmness and rebound) of eight experimental pillows and one presently marketed pillow. Because the company knew from previous studies that most people's attention span allowed for them to evaluate at most three pillows at a given time, it decided to employ the design shown in Table 19.16.

After the pillow types were randomly assigned the letters from A to I, tables were prepared with the appropriate pillow types assigned to each table. Each pillow was sealed in an identical white pillowcase and hence could not be distinguished from the others by color. The only marking on the pillowcase was a four-digit number, which provided the investigators with an identification code. With all tables in place, the 12 potential consumers were randomly assigned to a table to compare the three pillows. The consumers were to rate each pillow with a comfort score, based on a 1- to 100-point scale (a higher score indicates greater comfort). The scores for each pillow are recorded in Table 19.16 (letters identify the pillow type, with A being the presently marketed pillow).

TABLE 19.16
Comfort scores for Example 19.9

Block (consumers)	Treatment (pillow)			Block Total	Block Mean
1	A 59	B 26	C 38	123	41
2	D 85	E 92	F 69	246	82
3	G 74	H 52	I 27	153	51
4	A 63	D 70	G 68	201	67
5	B 26	E 98	H 59	183	61
6	C 31	F 60	I 35	126	42
7	A 62	E 85	I 30	177	59
8	B 23	F 73	G 75	171	57
9	C 49	D 74	H 51	174	58
10	A 52	F 76	H 43	171	57
11	B 18	D 79	I 41	138	46
12	C 42	E 84	G 81	207	69
				2,070	57.5

Verify that the design used is a BIB design. Use the formulas of this section to perform an analysis of variance. Use $\alpha = .05$ to test for a difference in mean comfort scores among the nine pillow types.

Solution We need to verify that all the conditions required for a BIB design have been satisfied. We note that there were nine treatments (pillows), 12 blocks (consumers), and three observations per block (pillows per consumer) and that each pillow was rated by four consumers, with a consumer rating, at most, one pillow of each type. That is, $t = 9$, $b = 12$, $k = 3$, and $r = 4$, which yield $n = (9)(4) = (12)(3) = 36$.

We next compute $\lambda = r(k-1)/(t-1) = 4(3-1)/(9-1) = 1$. That is, each pair of pillows was rated by exactly one consumer. We confirm this by examining

Table 19.16. Thus, the design used in the study was a BIB design. For an analysis using the formulas given in this section, it is convenient to construct a table of totals and means, as shown in Table 19.17.

TABLE 19.17
Totals for the data of Table 19.16

Treatment	$y_{i.}$	$B_{(i)}$	$ky_{i.} - B_{(i)}$
A	236	672	36
B	93	615	−336
C	160	630	−150
D	308	759	165
E	359	813	264
F	278	714	120
G	298	732	162
H	205	681	−66
I	133	594	−195
Total	2,070		0

To illustrate the values in Table 19.17, let us consider the elements for treatment A:

$$y_{1.} = \text{sum of values for treatment A} = 59 + 63 + 62 + 52 = 236$$
$$B_{(1)} = \text{sum of block totals for blocks containing A} = 123 + 201 + 177 + 171 = 672$$
$$ky_i - B_{(i)} = (3)(236) - 672 = 36$$

To compute the sums of squares, using the values in Tables 19.16 and 19.17, we have

$$\text{SST}_{\text{adj}} = \frac{(t-1)}{nk(k-1)} \sum_i (\text{k}y_{i.} - B_{(i)})^2 = \frac{(9-1)(316{,}638)}{(36)(3)(3-1)} = 11{,}727.33$$

Similarly, using the block means from Table 19.16, we obtain

$$\text{SSB} = k \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 = 3\{(41 - 57.5)^2 + \cdots + (69 - 57.5)^2\} = 4{,}575$$

Using the values from Table 19.16, we obtain the total sum of squares

$$\text{TSS} = \sum_{ij} (y_{ij} - \bar{y}_{..})^2 = \{(59 - 57.5)^2 + \cdots + (81 - 57.5)^2\} = 16{,}861$$

and the sum of squares for error

$$\text{SSE} = \text{TSS} - \text{SST}_{\text{adj}} - \text{SSB} = 16{,}861 - 11{,}727.33 - 4{,}575 = 558.67$$

The analysis of variance table for testing for differences in the mean comfort values among the nine types of pillows is shown in Table 19.18. Since the computed value of F, 41.98, exceeds the table value, 2.59, for $\text{df}_1 = 8$, $\text{df}_2 = 16$, and $\alpha = .05$, we conclude that there are significant (p-value $< .0001$) differences in the mean comfort ratings among the nine types of pillows.

TABLE 19.18
AOV table for the data of Example 19.9

Source	SS	df	MS	F	p-value
Consumer	4,575	11	415.91	—	—
Treatment	11,727.33	8	1,465.92	41.98	.0001
Error	558.67	16	34.92	—	—
Total	16,861	35	—	—	—

comparison among treatment means

Following the observation of a significant F test concerning differences among treatment means, we would then make **comparisons among treatment means**. To do this, we make use of the following notation: $\hat{\mu}_i$, an estimate of the mean for treatment i, given by

$$\hat{\mu}_i = \bar{y}_{..} + \frac{ky_{i.} - B_{(i)}}{t\lambda}$$

where $\bar{y}_{..}$ is the overall sample mean. An estimate of the difference between two treatment means, i and i', is then

$$\hat{\mu}_i - \hat{\mu}_{i'.} = \frac{[ky_{i.} - B_{(i)}] - [ky_{i'.} - B_{(i')}]}{t\lambda}$$

The Tukey-Kramer W for comparing any pair of treatment means is

$$W = \frac{q_\alpha(t, v)}{\sqrt{2}}\sqrt{\frac{2k\text{MSE}}{t\lambda}}$$

EXAMPLE 19.10

Compute the estimated treatment means and determine all pairwise differences, using $\alpha = .05$, for the data in Example 19.9.

Solution For the BIB design of Example 19.9, $\bar{y}_{..} = 57.5$, $t = 9$, and $\lambda = 1$. Thus, using the $ky_{i.} - B_{(i)}$ column in Table 19.17, we compute the estimated treatment means shown in Table 19.19 with

$$\hat{\mu}_i = \bar{y}_{..} + \frac{ky_{i.} - B_{(i)}}{t\lambda} = 57.5 + \frac{ky_{i.} - B_{(i)}}{(9)(1)}$$

TABLE 19.19
Estimated treatment means

Treatment	$\bar{y}_{i.}$	$ky_{i.} - B_{(i)}$	$\hat{\mu}_i$
A	59.00	36	61.50
B	23.25	−336	20.17
C	40.00	−150	40.83
D	77.00	165	75.83
E	89.75	264	86.83
F	69.50	120	70.83
G	74.50	162	75.50
H	51.25	−66	50.17
I	33.25	−195	35.83

Note that when comparing the raw treatment means, $\bar{y}_{i.}$, to the least-squares estimated means, $\hat{\mu}_i$, some of the raw means are increased, whereas some are decreased depending on the relative sizes of the block totals in which the treatment appears.

Using MSE = 34.92, based on $\text{df}_{\text{Error}} = 16$, we obtain

$$W = \frac{q_{.05}(9, 16)}{\sqrt{2}}\sqrt{\frac{2k\text{MSE}}{t\lambda}} = 3.56 \quad \sqrt{\frac{2(3)(34.92)}{(9)(1)}} = 17.18$$

The nine least-squares estimated treatment means are arranged in ascending order, with a summary of the significant results. Those treatments with a common

letter are not significantly different from each other, using the value of W to declare pairs significantly different.

B	I	C	H	A	F	G	D	E
20.17	35.83	40.83	50.17	61.50	70.83	75.50	75.83	86.83
a	ab	b	bc	cd	de	de	de	e

Alternatively, the computation of the adjusted sum of squares for treatments and the corresponding F test for testing differences in the treatment means can be accomplished by fitting two models. First, fit a full model with both block and treatment effects to obtain SSE_1. Next, fit a reduced model without treatments effects to obtain SSE_2. The adjusted sum of squares for treatments, SST_{Adj}, is then obtained by

$$SST_{Adj} = SSE_2 - SSE_1$$

with $df_{Trt} = df_{E2} - df_{E1}$. The F test for treatment effects is then $F = MST_{Adj}/MSE_1$.

EXAMPLE 19.11

Refer to Example 19.9. Use the following output to compute the sums of squares for treatments and error. Compare these values to the values computed using the estimated missing value formulas in Example 19.9.

```
SAS Output from The GLM Procedure

Class        Levels     Values
C                12     C1 C10 C11 C12 C2 C3 C4 C5 C6 C7 C8 C9
P                 9     P1 P2 P3 P4 P5 P6 P7 P8 P9


Number of Observations Read        108
Number of Observations Used         36

Model 1: Full Model

Dependent Variable: Y=Rating

                                  Sum of
Source                DF         Squares   Mean Square   F Value   Pr > F
Model                 19     16302.33333     858.01754     24.57   <.0001
Error                 16       558.66667      34.91667
Corrected Total       35     16861.00000

Source                DF     Type III SS   Mean Square   F Value   Pr > F
Consumer              11       454.33333      41.30303      1.18   0.3694
Pillow                 8     11727.33333    1465.91667     41.98   <.0001


Model 2: Reduced Model:

Dependent Variable: Y=Ratings

                                  Sum of
Source                DF         Squares   Mean Square   F Value   Pr > F
Model                 11      4575.00000     415.90909      0.81   0.6284
Error                 24     12286.00000     511.91667
Corrected Total       35     16861.00000

Source                DF     Type III SS   Mean Square   F Value   Pr > F
Consumer              11     4575.000000    415.909091      0.81   0.6284
```

From the full model, we obtain SSE = 558.67 with df = 16.

Using the full model and the reduced model, we obtain the adjusted sum of squares for the treatment (pillow):

$$SST_{adj} = SSE_{reduced} - SSE_{full} = 12{,}286 - 558.67 = 11{,}727.33$$

with df = 24 − 16 = 8.

Using the reduced model, we obtain the unadjusted sum of squares for blocks (consumers):

$$SSB = 4{,}575$$

These are the same values that we obtained in Example 19.9 using the estimated missing value formulas. ■

19.5 RESEARCH STUDY: Evaluation of the Consistency of Property Assessors

As was described in Section 19.1, there were a large number of complaints concerning the assessed valuation of residential homes by residents in a county located in a southwestern state. A group of property owners informed the county manager that there was wide variation in residential property valuations depending on which county property assessor determined the property's value. There are numerous assessors who determine the value of residential property for the purposes of computing property taxes due from each property owner in the county. The county manager designed a study to determine whether the assessors differ systematically in their determinations of property values.

The objective of the study was to determine whether the county assessors provided a consistent valuation of residential property values. The factors in the study were the blocking factor, 16 residential properties, and the treatment factor, 16 county property assessors. The treatment effects are random because the assessors were randomly selected from the population of county assessors and the county manager was interested in the results not only for the 16 assessors in the study but also for all county assessors.

The assessed valuations provided by the 16 assessors (in thousands of dollars) are presented in Table 19.20.

The design was an incomplete block design because each treatment (assessor) was observed in only 6 of the 16 blocks (properties). We will next verify that the design was a BIB design.

First, we identify the parameters in a BIB:

$$t = 16 \qquad r = 6 \qquad b = 16 \qquad k = 6$$

This would require that $n = (16)(6) = 96$ observations and $\lambda = 6(6 - 1)/(16 - 1) = 2$. From this, we would conclude that for the study to be a BIB design, it is necessary for every pair of assessors to valuate two of the same properties, each assessor must valuate 6 of the 16 properties, and we have a total of 96 valuations. An examination of the data reveals that all these conditions have been satisfied. We next fit the models necessary for an evaluation of the data. The model for relating the variation in valuations to assessor effects, property effects, and all other sources is given by

$$\text{Full model: } y_{ijg} = \mu + \tau_i + \beta_j + \varepsilon_{ijg}$$

TABLE 19.20 Property assessments (in thousands of dollars) by 16 county assessors

	Assessor															
Property	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1				125	120					112		115	118		110	
2				126		118		110		128	125			125		
3	110								125	118	138	110				126
4		131	150	157			125				150	156				
5	150	154	152		125			157		139						
6		138			118	110			120		124		129			
7	134				144	146	130					130		145		
8	157	159		150		134									120	158
9			156		155						150			138	124	156
10			156			128		155				153	155			122
11	155		158	157					142				123	155		
12		118					110			113			118	125		111
13			152			111	150		112	128					130	
14	115						112	110			135		130		128	
15		115						110	145			135		124	120	
16				157	120		150	135	120							132

where μ is the overall mean valuation across all assessors, τ_i is the random effect on the valuation due to assessor i, β_j is the random effect on the valuation due to property j, and ε_{ijg} represents the random effect of all other sources of variation on the valuation. Next, we fit the reduced models. First is the model without the assessor effect:

$$\text{Reduced model I: } y_{ijg} = \mu + \beta_j + \varepsilon_{ijg}$$

From this model, we obtain the adjusted sum of squares for assessors. Next, we fit the model without the property effect:

$$\text{Reduced model II: } y_{ijg} = \mu + \tau_i + \varepsilon_{ijg}$$

From this model, we obtain the adjusted sum of squares for properties.

The computer output given here provides us with the sums of squares for error from the three fitted models, SSE_{full}, $\text{SSE}_{\text{red I}}$, and $\text{SSE}_{\text{red II}}$.

```
General Linear Models: FULL MODEL

Dependent Variable: VALUATION

Source                  DF    Sum of Squares    F Value    Pr > F
Model                   30     16976.0919219       4.51    0.0001
Error                   65      8161.2414114
Corrected Total         95     25137.3333333

Source                  DF       Type III SS    F Value    Pr > F
ASR                     15      3759.0919219       2.00    0.0291
P                       15     10343.8800172       5.49    0.0001
```

```
General Linear Models: REDUCED MODEL WITHOUT TREATMENT VARIABLE (ASSESSOR)

Dependent Variable: VAL

Source                  DF    Sum of Squares    F Value    Pr > F
Model                   15     13217.0000000       5.91    0.0001
Error                   80     11920.3333333
Corrected Total         95     25137.3333333

Source                  DF       Type III SS    F Value    Pr > F
P                       15     13217.0000000       5.91    0.0001


General Linear Models: REDUCED MODEL WITHOUT BLOCK VARIABLE (PROPERTY)

Source                  DF    Sum of Squares    F Value    Pr > F
Model                   15      6632.21190476       1.91    0.0339
Error                   80     18505.12142857
Corrected Total         95      25137.3333333

Source                  DF       Type III SS    F Value    Pr > F
ASR                     15      6632.21190476       1.91    0.0339
```

The test for statistically significant differences in the mean valuations due to assessor differences is obtained as follows:

$$\text{SST}_{\text{adj}} = \text{SSE}_{\text{red I}} - \text{SSE}_{\text{full}} = 11{,}920.33 - 8{,}161.24 = 3{,}759.09$$

with $\text{df}_{\text{trt}} = \text{df}_{\text{Ered I}} - \text{df}_{\text{Efull}} = 80 - 65 = 15$. We can then test whether there is a significant variation in the valuations due to differences in the assessors. Since assessor is a random source of variation, we want to test

$$H_0\colon \sigma_\tau^2 = 0 \quad \text{versus} \quad H_a\colon \sigma_\tau^2 \neq 0$$

We compute the value of the test statistic

$$F = \frac{\text{SST}_{\text{adj}}/\text{df}_{\text{trt}}}{\text{SSE}_{\text{full}}/\text{df}_{\text{Efull}}} = \frac{3{,}759.09/15}{8{,}161.24/65} = 2.00$$

with p-value = .0291. We can compare the F-value to the tabled .05 percentile from an F distribution with $\text{df}_1 = 15$ and $\text{df}_2 = 65$, 1.82 and conclude that there is significant (p-value = .0291) variation due to the differences in the assessors. Similarly, we obtain the adjusted sum of squares due to the differences in the properties.

$$\text{SSB}_{\text{adj}} = \text{SSE}_{\text{red II}} - \text{SSE}_{\text{full}} = 18{,}505.12 - 8{,}161.24 = 10{,}343.88$$

with $\text{df}_{\text{block}} = \text{df}_{\text{Ered II}} - \text{df}_{\text{Efull}} = 80 - 65 = 15$. We can summarize our findings in an AOV table shown in Table 19.21.

TABLE 19.21
AOV table for research study

Source	df	SS	EMS	F	p-value
Property	15	10,343.88	$\sigma_\varepsilon^2 + 5.33\sigma_\beta^2$	—	—
Assessor	15	3,759.09	$\sigma_\varepsilon^2 + 5.33\sigma_\tau^2$	2.00	.0291
Error	65	8,161.24	σ_ε^2	—	—

Note that the multipliers for the variances from property and assessor effects are not 16, as they would be in a randomized complete design. Because of the incompleteness of the design, we have the following values for the expected mean squares:

$$\text{Expected mean square for blocks:} \quad \text{EMS}_{\text{block}} = \sigma_\varepsilon^2 + \frac{bk - t}{b - 1}\sigma_\beta^2$$

and

$$\text{Expected mean square for treatment:} \quad \text{EMS}_{\text{trt}} = \sigma_\varepsilon^2 + \frac{\lambda t}{k}\sigma_\tau^2$$

From Table 19.21, we can obtain the following estimates of the variance components:

$$\hat{\sigma}_\varepsilon^2 = 8{,}161.24/65 = 125.56$$

$$\hat{\sigma}_\beta^2 = (10{,}343.88/15 - 125.56)/5.33 = 105.82$$

$$\hat{\sigma}_\tau^2 = (3{,}759.09/15 - 125.56)/5.33 = 23.46$$

Thus, we have the proportional allocations of the total variability in the valuations, as shown in Table 19.22.

Although we found that there was significant (p-value = .0291) variability due to the assessors, less than 10% of the variability in the assessed valuations of the properties was due to assessors. Thus, we have determined that the assessors are reasonably consistent in their valuations of midpriced residental properties in the county.

Reporting Conclusions The report from the county staff to the county manager should include the following items.

1. Statement of objectives of study
2. Description of study design, how the properties used in the study were selected, how the assessors were selected, and the manner in which the valuations were conducted
3. Discussion of the relevance of the conclusions of this study to valuations throughout the county
4. Numerical and graphical representations of the data
5. Description of all inference methodologies:
   - Statement of research hypotheses
   - Model that represents experimental conditions
   - Verification of model conditions
   - AOV table, including p-values
6. Discussion of results and conclusions
7. Interpretation of findings relative to residential complaints about the biases in property valuations
8. Listing of data

TABLE 19.22
Allocation of total variance to sources

Source of Variation	Estimated Variance	Proportion of Total Variation (%)
Properties	105.82	41.5
Assessors	23.46	9.2
Exp. Error	125.56	49.3
Total	254.84	100

19.6 Summary and Key Formulas

In this chapter, we discussed the analysis of variance for some unbalanced designs, beginning with a discussion of the analysis for a randomized block design with one missing observation. Two possible analyses were proposed. The first required that we estimate the missing value and then proceed with the usual formulas developed in Chapter 15. Although estimating a single missing value is quite easy to do, the procedure becomes more difficult when there is more than one missing value. The second procedure, that of fitting complete and reduced models to obtain adjusted sums of squares, can be used for one or more missing observations.

With the Latin square design, we again showed how to estimate a single missing observation and proceed with the usual analysis. However, as with the randomized block design, the method of analysis by fitting complete and reduced models is more appropriate when there is more than one missing value.

Finally, we considered another class of unbalanced designs, incomplete block designs. The particular designs that we discussed were incomplete randomized block designs in which not all treatments appear in each block. These incomplete block designs retain a certain amount of balance because all pairs of treatments appear together in a block the same number of times. We illustrated the analysis for balanced incomplete block designs using appropriate formulas. The method of analysis for BIB designs can be accomplished by fitting full and reduced models as was done in the case of missing values in the randomized block design and Latin square design.

Key Formulas

1. Missing observation, y_{kh}, in a randomized block design

   a. $$\hat{y}_{kh} = \frac{ty_{k.} + by_{.h} - y_{..}}{(t-1)(b-1)}$$

   b. Bias correction for sum of squares for treatment

   $$\text{Bias} = \frac{(y_{.h} - (t-1)\hat{y}_{kh})^2}{t(t-1)}$$

   The corrected treatment sum of squares is then $\text{SST}_C = \text{SST} - \text{bias}$.

2. Tukey-Kramer W for a randomized block design

   a. For any pair of treatments with no missing value

   $$W = q_\alpha(t, v)\sqrt{\frac{\text{MSE}}{b}}$$

   b. Between the treatment with a missing value and any other treatment

   $$W^* = \frac{q_\alpha(t, v)}{\sqrt{2}}\sqrt{\text{MSE}\left(\frac{2}{b} + \frac{t}{b(b-1)(t-1)}\right)}$$

3. Equalities for randomized block design

   $$\text{SSB} = \text{TSS} - \text{SST}_{\text{adj}} - \text{SSE}$$
   $$\text{SST} = \text{TSS} - \text{SSB}_{\text{adj}} - \text{SSE}$$

4. Missing observation, y_{ghm}, in a Latin square design

   a. $$\hat{y}_{ghm} = \frac{t(y_{g..} + y_{.h.} + y_{..m}) - 2y_{...}}{(t-1)(t-2)}$$

b. Bias correction for sum of squares for treatment

$$\text{Bias} = \left(\frac{y_{...} - y_{g..} - y_{.h.} - (t-1)y_{..m}}{(t-1)(t-2)}\right)^2$$

The corrected treatment sum of squares is then $\text{SST}_C = \text{SST} - \text{bias}$.

5. Tukey-Kramer W for a Latin square design

a. For any pair of treatments with no missing value

$$W = q_\alpha(t, v)\sqrt{\frac{\text{MSE}}{t}}$$

b. Between the treatment with the missing value and any other treatment

$$W^* = \frac{q_\alpha(t, v)}{\sqrt{2}}\sqrt{\text{MSE}\left(\frac{2}{t} + \frac{1}{(t-1)(t-2)}\right)}$$

6. Sums of squares for an incomplete block design

$$\text{SST}_{\text{adj}} = \frac{t-1}{n(k)(k-1)}\sum_i (ky_{i.} - B_{(i)})^2$$

$$\text{SSE} = \text{TSS} - \text{SSB} - \text{SST}_{\text{adj}}$$

7. Pairwise comparisons of treatment means for an incomplete block design

$$\hat{\mu}_i - \hat{\mu}_{i'} = \frac{[ky_{i.} - B_{(i)}] - [ky_{i'.} - B_{(i')}]}{t\lambda}$$

$$W = \frac{q_\alpha(t, v)}{\sqrt{2}}\sqrt{\frac{2k\text{MSE}}{t\lambda}}$$

8. In a balanced incomplete block design

a. $n = rt = kb$.

b. $\lambda < r < b$.

c. $\lambda = r(k-1)/(t-1)$ must be an integer.

19.7 Exercises

19.2 A Randomized Block Design with One or More Missing Observations

Ag.

19.1 In Exercise 15.1, we described an experiment in which a horticulturist was investigating the effectiveness of five methods for the irrigation of blueberry shrubs. The methods are surface, trickle, center pivot, lateral move, and subirrigation. There are 10 blueberry farms available for the study representing a wide variety of types of soils, terrains, and wind gradients. The horticulturist wants to use each of the five methods of irrigation on all 10 farms to moderate the effect of the many extraneous sources of variation that may impact the blueberry yields. Each farm is divided into five plots, and the response variable will be the weight of the harvested fruit from each plot of blueberry shrubs. During the study, a problem occurred on the plot irrigated using the surface method on farm 1, and no yield was obtained. The yields in pounds of blueberries over a growing season are given here.

	Method of Irrigation				
Farm	**Surface**	**Trickle**	**Center Pivot**	**Lateral**	**Subirrigation**
1	*	248	391	423	350
2	636	382	434	461	370
3	591	348	492	504	460
4	603	366	468	580	452
5	649	258	457	449	343
6	512	321	406	464	340

(*continues*)

(*continued*)

	Method of Irrigation				
Farm	**Surface**	**Trickle**	**Center Pivot**	**Lateral**	**Subirrigation**
7	588	423	466	550	327
8	689	406	502	526	378
9	690	400	559	469	419
10	608	380	469	550	458

a. Estimate the yield value for the missing plot.
b. Analyze the data by replacing the missing value with the estimate obtained in part (a), and then perform an analysis of variance using the formulas for a randomized block design with no missing observations.
c. Is there a significant difference in the mean yields for the different methods of irrigation? Use $\alpha = 0.05$.

19.2 Refer to Exercise 19.1. Use the least significant difference criterion to identify which pairs of methods of irrigation have significantly different mean yields.

19.3 Refer to Exercise 19.1. Obtain the sums of squares for an AOV table by fitting complete and reduced models using a statistical software program. Compare your results with those in Exercise 19.1.

Edu.

19.4 The business office of a large university is in the process of selecting amongst the Postal Service and three private couriers as its sole delivery method for the university's responses to applications for admission. After consulting with the university's statistics department, it was decided that over the next month the following study would be conducted. Ten cities with at least 100 applicants would be selected for inclusion in the study. To each of these cities 100 standard packages would be sent by each of the four methods of delivery. The percentage of packages not delivered within 5 days was recorded for each method of delivery, yielding the following data. For four of the cities, at least one of the methods of delivery did not provide service, and, hence, there are missing data in these cells.

	City									
Method	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**
Ml	*	90.2	82.9	89.4	98.0	91.5	97.2	83.4	88.6	*
M2	87.1	99.5	92.0	91.4	99.2	91.5	97.6	88.7	92.7	97.6
M3	91.6	99.7	*	99.2	99.3	98.1	98.2	95.4	93.7	98.3
M4	95.5	99.9	93.8	98.9	99.4	98.6	*	94.1	93.1	99.3

a. Obtain the sums of squares for an AOV table by fitting complete and reduced models using a statistical software program.
b. Is there significant evidence of a difference in the four methods of delivery based on the percentage of packages delivered within 5 days?

19.5 Refer to Exercise 19.4. Use the Tukey-Kramer W procedure to identify which pairs of methods of delivery have significantly different mean percentages.

19.3 A Latin Square Design with Missing Data

Env.

19.6 Carbon monoxide (CO) emissions from automobiles can be influenced by the formulation of the gasoline that is used. Oxygenated fuels are used in northern states during the winter to decrease CO emissions. There are eight gasoline blends that are of interest to the researchers (B1—B8). Each of the eight blends will be placed in a car that will then be driven over a 50-mile route during which the total amount of CO emissions will be measured. There are large car-to-car differences in CO emissions, and there are large route-to-route differences in city driving (stop-and-go driving on city streets versus a freeway route). The researchers have eight cars and eight routes available to study the eight blends, with every blend observed in all eight cars, which will be driven over all eight routes. The following table contains the amount of CO emissions (grams) per mile by each vehicle, route, and blend. During the study, the device used to measure CO

emissions failed to function properly when vehicle V7 was driven over route R3 using blend B1. The research goal is to determine how the different blends impact the mean CO readings.

	Route							
Vehicle	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**
V1	B1 12.0	B2 11.2	B3 11.8	B4 10.0	B5 20.1	B6 18.7	B7 21.7	B8 30.2
V2	B2 10.1	B3 12.2	B4 12.1	B5 12.4	B6 18.4	B7 18.6	B8 22.3	B1 15.0
V3	B3 21.4	B4 24.2	B5 26.7	B6 23.3	B7 32.5	B8 34.1	B1 21.4	B2 27.7
V4	B4 15.4	B5 20.3	B6 17.5	B7 17.6	B8 25.3	B1 12.2	B2 12.4	B3 18.9
V5	B5 25.0	B6 24.4	B7 24.0	B8 26.5	B1 20.6	B2 19.6	B3 19.6	B4 27.3
V6	B6 18.9	B7 20.9	B8 25.2	B1 8.3	B2 15.6	B3 15.1	B4 17.4	B5 25.9
V7	B7 16.2	B8 18.2	B1 ***	B2 4.4	B3 10.2	B4 9.9	B5 12.7	B6 17.9
V8	B8 29.5	B1 21.3	B2 18.3	B3 16.1	B4 26.0	B5 26.4	B6 26.0	B7 35.0

a. Estimate the amount of CO emissions for vehicle V7 while driving over route R3 using blend B1.
b. Analyze the data by replacing the missing value with the estimate obtained in part (a), and then perform an analysis of variance using the formulas for a Latin square design with no missing observations.
c. Is there a significant difference in the mean CO emissions for the different blends? Use $\alpha = .05$.

19.7 Refer to Exercise 19.6. Use the Tukey-Kramer W to identify which pairs of blends have significantly different mean CO emissions.

19.8 Refer to Exercise 19.6. Obtain the sums of squares for an AOV table by fitting complete and reduced models using a statistical software program. Compare your results with those in Exercise 19.7.

19.9 Refer to Exercise 19.6. Suppose upon examining the data logs from the study the researchers determined that the CO emissions monitoring device was probably not functioning properly for the following two data values: vehicle V7 on route R4 using blend B2, y_{742}, and vehicle V6 on route R4 using blend B1, y_{641}. Reanalyze the data after deleting these two values. Do your conclusions about the differences in the eight blends change?

19.10 Refer to Exercise 19.9.
a. Identify vehicle and route as fixed or random effects.
b. How would you test for a significant effect due to vehicle?
c. How would you test for a significant effect due to route?

Sci.

19.11 A horticulturist is interested in examining the yield potential of three new varieties of asparagus. She designed a study to evaluate the three new varieties relative to the standard variety. There were 16 plots available on a large test field for the study, but the plots were not homogeneous in that there was a distinct sloping from north to south throughout the field. Also, a soil analysis revealed a discernible nitrogen gradient, which ran from west to east across the field. Therefore, the horticulturists decided to assign the varieties V1, V2, V3, and V4, with V1 being the standard variety, to the plots in a Latin square arrangement. The values for marketable yield per plot (in kg/ha) are given in the following table. Note that there is a missing yield for variety V4 in row 4 and column 1. This was due to a problem that occurred during one of the harvesting periods.

	Sloping			
Nitrogen	**S1**	**S2**	**S3**	**S4**
N1	V3 1,045.38	V1 807.69	V2 967.36	V4 1,084.23
N2	V1 821.40	V2 992.56	V4 992.47	V3 1,029.53
N3	V2 1,004.02	V4 1,091.23	V3 1,062.01	V1 836.53
N4	V4 *	V3 1,090.97	V1 893.32	V2 1,053.97

a. Estimate the amount of marketable yield for variety V4 planted in a plot with nitrogen level N4 and slope S1.
b. Analyze the data by replacing the missing value with the estimate obtained in part (a), and then perform an analysis of variance using the formulas for a Latin square design with no missing observations.
c. Is there a significant difference in the mean marketable yields for the four varieties? Use $\alpha = 0.05$.

19.12 Refer to Exercise 19.11. Use the Tukey-Kramer W to identify which pairs of varieties have significantly different mean marketable yields.

19.13 Refer to Exercise 19.11. Obtain the sums of squares for an AOV table by fitting complete and reduced models using a statistical software program. Compare your results with those in Exercise 19.12.

19.14 Refer to Exercise 19.11.
a. Identify nitrogen level and slope level as either fixed or random effects.
b. How would you test for a significant difference in the mean marketable yields due to differences in nitrogen levels?
c. How would you test for a significant difference in the mean marketable yields due to differences in the amount of slope in the plots?

19.4 Balanced Incomplete Block (BIB) Designs

19.15 An incomplete block design consisted of five blocks (B1, B2, B3, B4, and B5) and five treatments (T1, T2, T3, T4, and T5). The treatments were randomly assigned to the blocks in the following manner.

Block	Treatments			
B1	T5	T1	T4	T3
B2	T2	T5	T4	T3
B3	T2	T1	T4	T3
B4	T2	T5	T1	T4
B5	T2	T5	T1	T3

a. What are the values of the design parameters: t, k, b, and r?
b. What is the value of λ for this design?
c. Is the incomplete block design balanced? Justify your answer.

19.16 An incomplete block design consisted of six blocks (B1, B2, B3, B4, B5, and B6) and six treatments (T1, T2, T3, T4, T5, and T6). The treatments were randomly assigned to the blocks in the following manner.

Block	Treatments		
B1	T5	T6	T1
B2	T3	T4	T1
B3	T5	T2	T4
B4	T2	T6	T1
B5	T3	T4	T6
B6	T5	T2	T3

a. What are the values of the design parameters: t, k, b, and r?
b. What is the value of λ for this design?
c. Is the incomplete block design balanced? Justify your answer.

Sci. **19.17** A study of the difference in the effects of six newly created diets on the weight gain of young rabbits is proposed. Because weight varies considerably amongst young rabbits, it is proposed to block the experiment based on litters. There are 10 litters of rabbits available for the study, but they are of varying sizes. The minimum litter size is three. Therefore, only three of the six diets can be observed in any particular litter. A balanced incomplete block design was proposed for this situation. The researcher conducted the study and obtained the following weight gains.

	Diet							
Litter	**1**	**2**	**3**	**4**	**5**	**6**	**Litter Total**	**Litter Mean**
1		32.6	35.2			42.2	110.0	36.67
2	40.1	38.1	40.9				119.1	40.43
3			34.6	37.5		34.3	106.4	39.70
4	44.9		43.9		40.8		129.6	35.47
5			40.9	37.3	32.0		110.2	43.20
6		37.3			40.5	42.8	120.6	36.73
7	45.2	40.6		37.9			123.7	40.20
8	44.0				38.5	51.9	134.4	41.23
9		30.6		27.5	20.6		78.7	44.80
10	37.3			42.3		41.7	121.3	26.23
Diet total	211.5	179.2	195.5	182.5	172.4	212.9	1,154.0	
Diet mean	42.3	35.84	39.1	36.5	34.48	42.58		38.47

Do the data provide significant evidence of a difference in mean weight gains amongst the six diets? Use the formulas given in this Section 19.4 to obtain your answers.

Sci. **19.18** Refer to Exercise 19.17. Use the Tukey-Kramer W to determine which pairs of diets have significantly different mean weight gains.

Sci. **19.19** Refer to Exercise 19.17. Analyze the data using a computer program. Is the analysis of variance table from the output of the computer program the same as your results in Exercise 19.18?

Sci. **19.20** Refer to Exercise 19.17. Test for a significant effect due to litter.

Supplementary Exercises

Env. **19.21** A petroleum company was interested in comparing the miles per gallon achieved by four different gasoline blends (I, II, III, and IV). Because there can be considerable variability due to differences in drivers and car models, these two extraneous sources of variability were included as blocking variables in the following Latin square design. Each driver drove each car model over a standard course with the assigned gasoline blend. However, when driver 3 was operating car model 4 using blend II gasoline, there was a malfunction of the car's carburator that invalidated the data. This malfunction was not discovered until well after the completion of the study, and, hence, the data could not be replaced. The miles per gallon data are given here.

	Car Model			
Driver	**1**	**2**	**3**	**4**
1	IV 15.5	II 33.9	III 13.2	I 29.1
2	II 16.3	III 26.6	I 19.4	IV 22.8
3	III 10.8	I 31.1	IV 17.1	II —
4	I 14.7	IV 34.0	II 19.7	III 21.6

a. Run an analysis of variance by estimating the missing value. Use $\alpha = .05$.
b. Make treatment comparisons by using the Tukey-Kramer W, with $\alpha = .05$.

Env. **19.22** Use the method of fitting complete and reduced models to obtain an analysis of variance for the data in Exercise 19.21.

Med. **19.23** A physician was interested in comparing the effects of six different antihistamines in persons extremely sensitive to antihistamine injections. To do this, a random sample of 10 allergy patients was selected from the physician's private practice, with treatments (antihistamines) assigned to each patient according to the experimental design shown in the following table. Each person then received injections of the assigned antihistamines in different sections of the right arm. The area of redness surrounding the point of injection was measured after a fixed period of time. The data are shown in the table.

Person	Treatments					
1	B	25	A	41	F	40
2	E	37	B	46	A	42
3	C	45	D	33	B	37
4	E	34	D	35	A	46
5	B	31	F	42	D	34
6	C	56	E	36	F	65
7	D	33	A	42	C	67
8	F	49	D	37	E	30
9	C	59	A	40	F	55
10	B	36	C	57	E	34

a. Identify the design.
b. Identify the characteristics of the design.
c. Run an analysis of variance. Use $\alpha = .05$.

Med. **19.24** Refer to Exercise 19.23. Use the Tukey-Kramer W for determining treatment differences, with $\alpha = .05$.

Psy. **19.25** The marketing research group of a corporation examined the public response to the introduction of a new TV game module by comparing weekly sales volumes (in $ thousand) for three different store chains in each of four geographic locations.

		Chain		
Geographic Area		1	2	3
N	W1	35	17	7
	W2	30	22	12
S	W1	42	30	22
	W2	48	28	19
E	W1	35	35	15
	W2	38	40	20
W	W1	22	43	28
	W2	26	48	23

a. Write an appropriate model (including an effect for weeks) and the sources of variability in an analysis of variance table.
b. How would your model change if we analyze the total 2-week sales data?
c. Run an analysis of variance on the 2-week sales data using formulas from Chapter 15. Use $\alpha = .05$.

Psy. **19.26** Refer to Exercise 19.25. Use the Tukey-Kramer W procedure to compare the different geographic areas by chain means. Use $\alpha = .05$.

Psy. **19.27** Refer to Exercise 19.26. Suppose that the week 1 data were not available in the north and east for chain 1, due to logistics problems that slowed the introduction of the product by a week.

a. Write an appropriate model.
b. Suggest a method for analyzing the data using available software.
c. Write model(s) for the procedure described in part (b).

H.R. **19.28** A foreign automobile manufacturer is spending hundreds of millions of dollars to construct a large manufacturing plant (about 70 acres under one roof) here in the United States. One of its objectives is to produce cars of high quality in the United States using U.S. workers. One part of the massive orientation program for new employees is to send about 20% of them to the home country for additional training. One measure of the worth of this additional training is whether the product quality is better on assembly lines where 20% of the employees have had the homeland orientation and have been able to share it with their fellow employees. Data from six assembly lines (three with the additional orientation) are shown here. To measure defects, two different inspectors examined each of two cars chosen at random from each of the assembly lines. Use these data to answer the following questions.

Additional Training			No Additional Training		
	Inspector			Inspector	
Assembly Line	1	2	Assembly Line	1	2
1	6	6	4	8	7
	3	4		5	5
2	4	3	5	10	9
	2	2		4	4
3	2	3	6	15	13
	1	1		7	6

a. Suggest an appropriate dependent variable.
b. Write a model for this experimental situation, and identify all terms.
c. Fill out the sources and degrees of freedom for an AOV table.

19.29 Refer to the conditions of Exercise 19.28.

a. Suggest a method to analyze these data.
b. Does the training produce fewer defects?
c. Can you suggest any plots that might be helpful in interpreting the data?

H.R. **19.30** Refer to Exercise 19.28. Suppose that inspector 2 was unable to evaluate the second car from assembly line 4 and that inspector 1 missed car 1 from assembly line 3.

a. Does the model change?
b. Suggest a method for analyzing the data.

Bus. **19.31** The state real estate commission is mandated to provide an examination that ensures a person passing the exam will have a minimum level of competence. This provides protection for the members of the public in their dealing with real estate firms. The state regulatory agency is responsible for establishing the acceptable level of safe practice and for determining whether an individual meets that standard. The real estate board has received several complaints about the grading of the essay questions on the exams. The board's staff designs a study to evaluate their current testing procedure by evaluating the differences in the grading of the essay questions on the real estate exam. The study included 25 real estate exam graders and a random sample of 30 exams taken during the past year. Because the grading of the exams is very time consuming, each grader was assigned 6 exams to score, with the scores given in the following table. The number in parenthesis is the identifer for the grader.

Exam	Score					Exam	Score				
1	70(1)	65(8)	61(15)	66(17)	66(24)	16	52(1)	54(6)	55(11)	62(16)	54(21)
2	84(2)	82(9)	86(11)	85(18)	86(25)	17	56(2)	51(7)	51(12)	52(17)	57(22)
3	72(3)	85(10)	77(12)	82(19)	79(21)	18	55(3)	60(8)	61(13)	59(18)	60(23)
4	85(4)	75(6)	78(13)	82(20)	83(22)	19	88(4)	76(9)	74(14)	77(19)	77(24)
5	58(5)	64(7)	58(14)	57(16)	58(23)	20	65(5)	68(10)	77(15)	72(20)	74(25)
6	66(1)	71(7)	73(13)	70(19)	70(25)	21	79(1)	77(10)	79(14)	77(18)	77(22)
7	73(2)	67(8)	63(14)	70(20)	66(21)	22	70(2)	66(6)	66(15)	63(19)	62(23)
8	58(3)	70(9)	69(15)	61(16)	71(22)	23	48(3)	49(7)	50(11)	51(20)	48(24)
9	95(4)	84(10)	88(11)	85(17)	87(23)	24	75(4)	64(8)	65(12)	75(16)	68(25)
10	47(5)	47(6)	51(12)	49(18)	56(24)	25	79(5)	77(9)	83(13)	81(17)	79(21)
11	60(1)	59(2)	51(3)	64(4)	53(5)	26	61(1)	67(9)	65(12)	69(20)	68(23)
12	64(6)	69(7)	63(8)	63(9)	71(10)	27	78(2)	75(10)	72(13)	76(16)	75(24)
13	84(11)	85(12)	86(13)	85(14)	83(15)	28	67(3)	72(6)	76(14)	72(17)	75(25)
14	72(16)	76(17)	77(18)	74(19)	77(20)	29	84(4)	81(7)	77(15)	76(18)	79(21)
15	65(21)	73(22)	70(23)	71(24)	70(25)	30	81(5)	84(8)	81(11)	85(19)	84(22)

a. Describe by name the type of design used. Verify that the structural conditions of your selected design are satisfied in this study.

b. Is there a difference in the average scores of the graders? Justify your answer at the $\alpha = .05$ level.

c. Was it necessary to include the exam factor in the design and subsequent analysis of the data?

d. Using the residuals, do there appear to be any violations in the conditions needed to run tests of hypotheses in the analysis of variances?

e. Do you think that the board should be concerned with the differences in the graders' evaluations of the exams if a difference of four units in their scores is deemed to be an important difference?

Chem.

19.32 Functionalized styrenes are extremely useful building blocks for organic synthesis and for functional polymers. One of the most general syntheses of styrenes involves the combination of an aryl halide with a vinyl organometallic reagent under catalysis by palladium (Pd) complexes. A study was designed to evaluate the effect of different levels of Pd—0.01, 0.05, 0.1, 0.5, and 1.0 (mol%)—on the yield of vinylboronic acid. The reactions take place in a high-pressure chamber at a temperature of 135°C. There are only three pressure chambers available for a single run of the experimental conditions. The chemists were concerned about the substantial run-to-run variations in the yields produced by new setups of the experiment in the chambers. Thus, it was necessary to block on runs, but with only three chambers, it was not possible to include all five levels of Pd during each run. The yields of vinylboronic acid are given in the following table.

	Paladium Level (mol%)				
Run	0.01	0.05	0.1	0.5	1.0
1	66	78	*	92	*
2	69	*	*	100	100
3	*	86	99	*	100
4	*	*	81	95	100
5	*	79	*	100	98
6	80	*	93	91	*
7	73	73	94	*	*
8	81	*	90	*	97
9	84	70	*	*	99
10	*	84	91	97	*

a. Describe by name the type of design used. Verify that the structural conditions of your selected design are satisfied in this study.
b. Is there a difference in the average yields of the five levels of paladium? Justify your answer at the $\alpha = .05$ level.
c. Was it necessary to include the runs factor in the design and subsequent analysis of the data?
d. Using the residuals, do there appear to be any violations in the conditions needed to run tests of hypotheses in the analysis of variances?
e. Do the levels of paladium appear to produce an important difference in average yields if a difference of 4% in yields is considered important?

APPENDIX

Statistical Tables

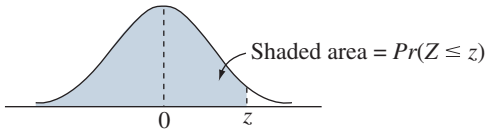


TABLE 1

Standard normal curve areas

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

z	Area
−3.50	.00023263
−4.00	.00003167
−4.50	.00000340
−5.00	.00000029
$-\infty$	.00000000

Source: Computed by M. Longnecker using the R function pnorm(z).

TABLE 1
(*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

z	Area
3.50	.99976737
4.00	.99996833
4.50	.99999660
5.00	.99999971
∞	1.0

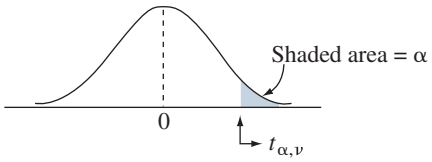


TABLE 2

Percentage points of Student's t distribution

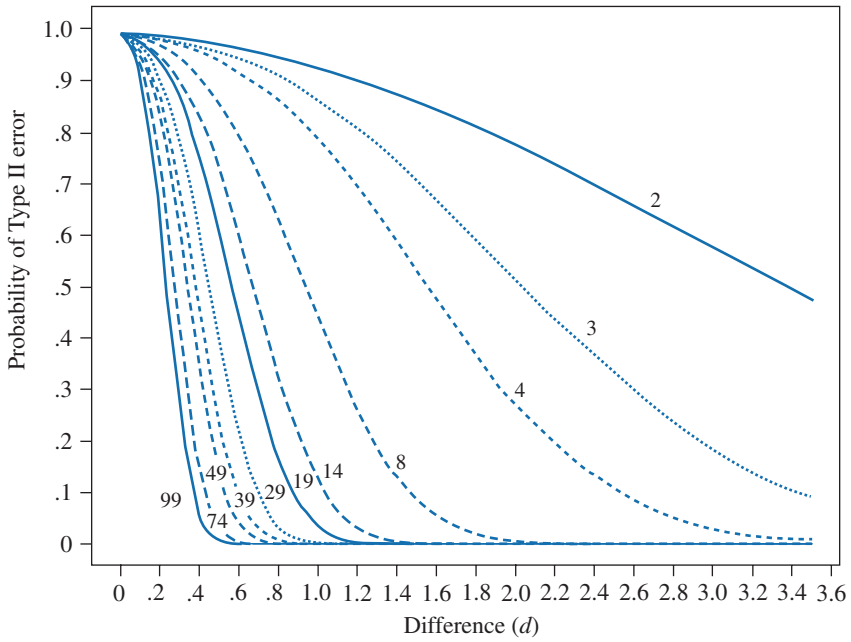
	Right-Tail Probability (α)								
df	**.40**	**.25**	**.10**	**.05**	**.025**	**.01**	**.005**	**.001**	**.0005**
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	.289	.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	.277	.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	.255	.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	.255	.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
inf.	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Source: Computed by M. Longnecker using the R function $qt(1 - \alpha, \text{df})$.

For level α two-tailed tests and $100(1 - \alpha)\%$ C.I.s use value in column headed by the number obtained by computing $\alpha/2$.

TABLE 3(a)

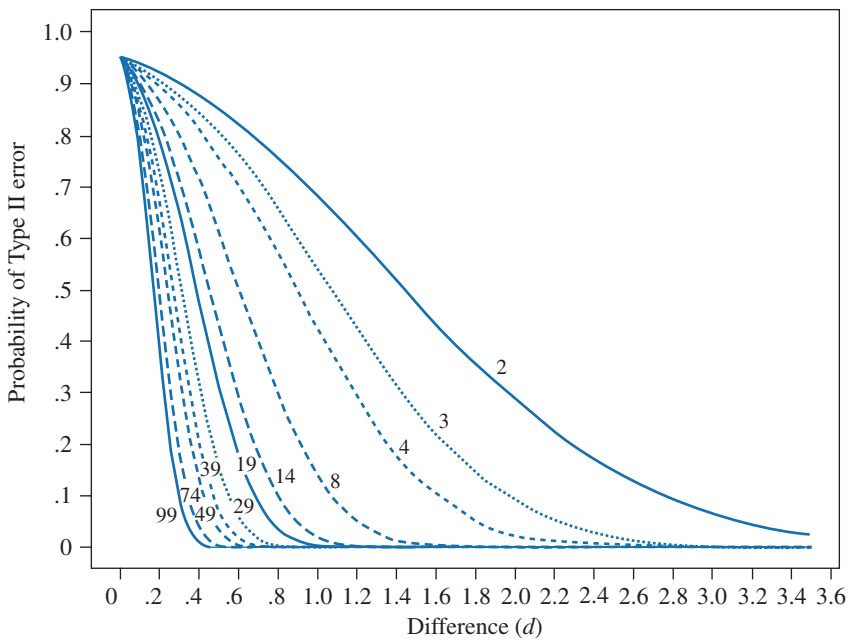
t Test probability of Type II error curves for $\alpha = .01$ (one-sided)



Source: Computed by M. Longnecker using SAS.

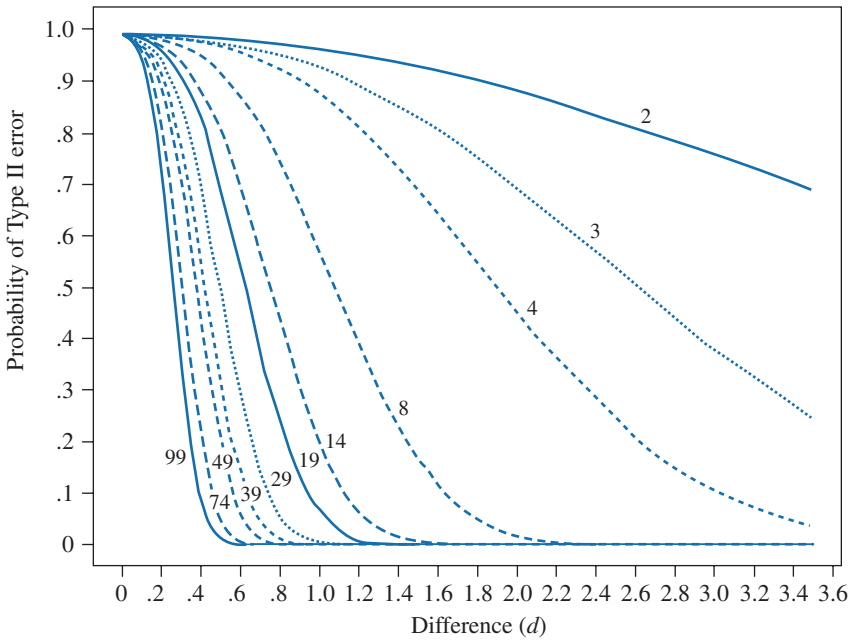
TABLE 3(b)

t Test probability of Type II error curves for $\alpha = .05$ (one-sided)



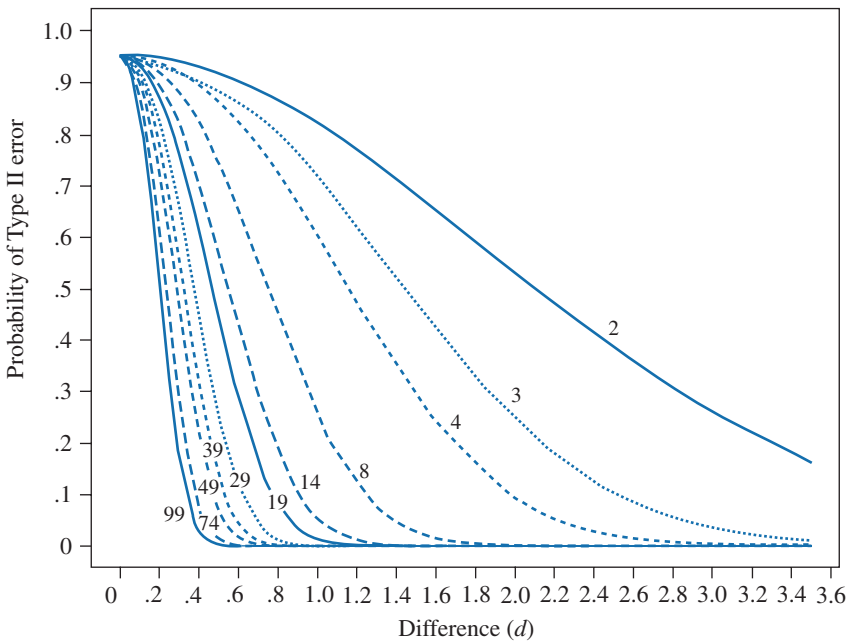
Source: Computed by M. Longnecker using SAS.

TABLE 3(c)
t Test probability of Type II error curves for $\alpha = .01$ (two-sided)



Source: Computed by M. Longnecker using SAS.

TABLE 3(d)
t Test probability of Type II error curves for $\alpha = .05$ (two-sided)



Source: Computed by M. Longnecker using SAS.

TABLE 4

Percentage points for confidence intervals on the median and the sign test: $C_{\alpha,n}$

$\alpha(2)$	.20	.10	.05	.02	.01	.005	.002	$\alpha(2)$	.20	.10	.05	.02	.01	.005	.002
$\alpha(1)$	.10	.05	.025	.01	.005	.0025	.001	$\alpha(1)$	.10	.05	.025	.01	.005	.0025	.001
n								***n***							
1	*	*	*	*	*	*	*	26	9	8	7	6	6	5	4
2	*	*	*	*	*	*	*	27	9	8	7	7	6	5	5
3	*	*	*	*	*	*	*	28	10	9	8	7	6	6	5
4	0	*	*	*	*	*	*	29	10	9	8	7	7	6	5
5	0	0	*	*	*	*	*	30	10	10	9	8	7	6	6
6	0	0	0	*	*	*	*	31	11	10	9	8	7	7	6
7	1	0	0	0	*	*	*	32	11	10	9	8	8	7	6
8	1	1	0	0	0	*	*	33	12	11	10	9	8	8	7
9	2	1	1	0	0	0	*	34	12	11	10	9	9	8	7
10	2	1	1	0	0	0	0	35	13	12	11	10	9	8	8
11	2	2	1	1	0	0	0	36	13	12	11	10	9	9	8
12	3	2	2	1	1	0	0	37	14	13	12	10	10	9	8
13	3	3	2	1	1	1	0	38	14	13	12	11	10	9	9
14	4	3	2	2	1	1	1	39	15	13	12	11	11	10	9
15	4	3	3	2	2	1	1	40	15	14	13	12	11	10	9
16	4	4	3	2	2	2	1	41	15	14	13	12	11	11	10
17	5	4	4	3	2	2	1	42	16	15	14	13	12	11	10
18	5	5	4	3	3	2	2	43	16	15	14	13	12	11	11
19	6	5	4	4	3	3	2	44	17	16	15	13	13	12	11
20	6	5	5	4	3	3	2	45	17	16	15	14	13	12	11
21	7	6	5	4	4	3	3	46	18	16	15	14	13	13	12
22	7	6	5	5	4	4	3	47	18	17	16	15	14	13	12
23	7	7	6	5	4	4	3	48	19	17	16	15	14	13	12
24	8	7	6	5	5	4	4	49	19	18	17	15	15	14	13
25	8	7	7	6	5	5	4	50	19	18	17	16	15	14	13

Note: An * means that no test or confidence interval of this level exists.

Source: Computed by M. Longnecker using the R function pbinom($c, n, .5$).

TABLE 5

Critical values for the Wilcoxon rank sum test: T_L and T_U.
Test statistic is rank sum associated with smaller sample (if equal sample sizes, either rank sum can be used).

a. $\alpha = .025$ one-tailed; $\alpha = .05$ two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131

b. $\alpha = .05$ one-tailed; $\alpha = .10$ two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42
5	7	20	13	27	19	36	20	40	22	43	24	46	25	50	26	54
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127

Source: From F. Wilcoxon and R. A. Wilcox, *Some Rapid Approximate Statistical Procedures* (Pearl River, NY: Lederle Laboratories, 1964), pp. 20–23. Reproduced with the permission of American Cyanamid Company.

TABLE 6
Critical values for the Wilcoxon signed-rank test [$n = 5(1)54$]

One-Sided	Two-Sided	$n = 5$	$n = 6$	$n = 7$	$n = 8$	$n = 9$
$p = .1$	$p = .2$	2	3	5	8	10
$p = .05$	$p = .1$	0	2	3	5	8
$p = .025$	$p = .05$		0	2	3	5
$p = .01$	$p = .02$			0	1	3
$p = .005$	$p = .01$				0	1
$p = .0025$	$p = .005$					0
$p = .001$	$p = .002$					

One-Sided	Two-Sided	$n = 10$	$n = 11$	$n = 12$	$n = 13$	$n = 14$
$p = .1$	$p = .2$	14	17	21	26	31
$p = .05$	$p = .1$	10	13	17	21	25
$p = .025$	$p = .05$	8	10	13	17	21
$p = .01$	$p = .02$	5	7	9	12	15
$p = .005$	$p = .01$	3	5	7	9	12
$p = .0025$	$p = .005$	1	3	5	7	9
$p = .001$	$p = .002$	0	1	2	4	6

One-Sided	Two-Sided	$n = 15$	$n = 16$	$n = 17$	$n = 18$	$n = 19$
$p = .1$	$p = .2$	36	42	48	55	62
$p = .05$	$p = .1$	30	35	41	47	53
$p = .025$	$p = .05$	25	29	34	40	46
$p = .01$	$p = .02$	19	23	27	32	37
$p = .005$	$p = .01$	15	19	23	27	32
$p = .0025$	$p = .005$	12	15	19	23	27
$p = .001$	$p = .002$	8	11	14	18	21

One-Sided	Two-Sided	$n = 20$	$n = 21$	$n = 22$	$n = 23$	$n = 24$
$p = .1$	$p = .2$	69	77	86	94	104
$p = .05$	$p = .1$	60	67	75	83	91
$p = .025$	$p = .05$	52	58	65	73	81
$p = .01$	$p = .02$	43	49	55	62	69
$p = .005$	$p = .01$	37	42	48	54	61
$p = .0025$	$p = .005$	32	37	42	48	54
$p = .001$	$p = .002$	26	30	35	40	45

One-Sided	Two-Sided	$n = 25$	$n = 26$	$n = 27$	$n = 28$	$n = 29$
$p = .1$	$p = .2$	113	124	134	145	157
$p = .05$	$p = .1$	100	110	119	130	140
$p = .025$	$p = .05$	89	98	107	116	126
$p = .01$	$p = .02$	76	84	92	101	110
$p = .005$	$p = .01$	68	75	83	91	100
$p = .0025$	$p = .005$	60	67	74	82	90
$p = .001$	$p = .002$	51	58	64	71	79

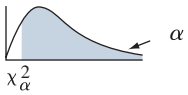
Source: Computed by P. J. Hildebrand.

TABLE 6
(*continued*)

One-Sided	Two-Sided	$n = 30$	$n = 31$	$n = 32$	$n = 33$	$n = 34$
$p = .1$	$p = .2$	169	181	194	207	221
$p = .05$	$p = .1$	151	163	175	187	200
$p = .025$	$p = .05$	137	147	159	170	182
$p = .01$	$p = .02$	120	130	140	151	162
$p = .005$	$p = .01$	109	118	128	138	148
$p = .0025$	$p = .005$	98	107	116	126	136
$p = .001$	$p = .002$	86	94	103	112	121
One-Sided	**Two-Sided**	$n = 35$	$n = 36$	$n = 37$	$n = 38$	$n = 39$
$p = .1$	$p = .2$	235	250	265	281	297
$p = .05$	$p = .1$	213	227	241	256	271
$p = .025$	$p = .05$	195	208	221	235	249
$p = .01$	$p = .02$	173	185	198	211	224
$p = .005$	$p = .01$	159	171	182	194	207
$p = .0025$	$p = .005$	146	157	168	180	192
$p = .001$	$p = .002$	131	141	151	162	173
One-Sided	**Two-Sided**	$n = 40$	$n = 41$	$n = 42$	$n = 43$	$n = 44$
$p = .1$	$p = .2$	313	330	348	365	384
$p = .05$	$p = .1$	286	302	319	336	353
$p = .025$	$p = .05$	264	279	294	310	327
$p = .01$	$p = .02$	238	252	266	281	296
$p = .005$	$p = .01$	220	233	247	261	276
$p = .0025$	$p = .005$	204	217	230	244	258
$p = .001$	$p = .002$	185	197	209	222	235
One-Sided	**Two-Sided**	$n = 45$	$n = 46$	$n = 47$	$n = 48$	$n = 49$
$p = .1$	$p = .2$	402	422	441	462	482
$p = .05$	$p = .1$	371	389	407	426	446
$p = .025$	$p = .05$	343	361	378	396	415
$p = .01$	$p = .02$	312	328	345	362	379
$p = .005$	$p = .01$	291	307	322	339	355
$p = .0025$	$p = .005$	272	287	302	318	334
$p = .001$	$p = .002$	249	263	277	292	307
One-Sided	**Two-Sided**	$n = 50$	$n = 51$	$n = 52$	$n = 53$	$n = 54$
$p = .1$	$p = .2$	503	525	547	569	592
$p = .05$	$p = .1$	466	486	507	529	550
$p = .025$	$p = .05$	434	453	473	494	514
$p = .01$	$p = .02$	397	416	434	454	473
$p = .005$	$p = .01$	373	390	408	427	445
$p = .0025$	$p = .005$	350	367	384	402	420
$p = .001$	$p = .002$	323	339	355	372	389

TABLE 7

Percentage points of the chi-square distribution



	Right-Tail Probability (α)					
df	**.999**	**.995**	**.99**	**.975**	**.95**	**.90**
1	.000002	.000039	.000157	.000982	.003932	.01579
2	.002001	.01003	.02010	.05064	.1026	.2107
3	.02430	.07172	.1148	.2158	.3518	.5844
4	.09080	.2070	.2971	.4844	.7107	1.064
5	.2102	.4117	.5543	.8312	1.145	1.610
6	.3811	.6757	.8721	1.237	1.635	2.204
7	.5985	.9893	1.239	1.690	2.167	2.833
8	.8571	1.344	1.646	2.180	2.733	3.490
9	1.152	1.735	2.088	2.700	3.325	4.168
10	1.479	2.156	2.558	3.247	3.940	4.865
11	1.834	2.603	3.053	3.816	4.575	5.578
12	2.214	3.074	3.571	4.404	5.226	6.304
13	2.617	3.565	4.107	5.009	5.892	7.042
14	3.041	4.075	4.660	5.629	6.571	7.790
15	3.483	4.601	5.229	6.262	7.261	8.547
16	3.942	5.142	5.812	6.908	7.962	9.312
17	4.416	5.697	6.408	7.564	8.672	10.09
18	4.905	6.265	7.015	8.231	9.390	10.86
19	5.407	6.844	7.633	8.907	10.12	11.65
20	5.921	7.434	8.260	9.591	10.85	12.44
21	6.447	8.034	8.897	10.28	11.59	13.24
22	6.983	8.643	9.542	10.98	12.34	14.04
23	7.529	9.260	10.20	11.69	13.09	14.85
24	8.085	9.886	10.86	12.40	13.85	15.66
25	8.649	10.52	11.52	13.12	14.61	16.47
26	9.222	11.16	12.20	13.84	15.38	17.29
27	9.803	11.81	12.88	14.57	16.15	18.11
28	10.39	12.46	13.56	15.31	16.93	18.94
29	10.99	13.12	14.26	16.05	17.71	19.77
30	11.59	13.79	14.95	16.79	18.49	20.60
40	17.92	20.71	22.16	24.43	26.51	29.05
50	24.67	27.99	29.71	32.36	34.76	37.69
60	31.74	35.53	37.48	40.48	43.19	46.46
70	39.04	43.28	45.44	48.76	51.74	55.33
80	46.52	51.17	53.54	57.15	60.39	64.28
90	54.16	59.20	61.75	65.65	69.13	73.29
100	61.92	67.33	70.06	74.22	77.93	82.36
120	77.76	83.85	86.92	91.57	95.70	100.62
240	177.95	187.32	191.99	198.98	205.14	212.39

Source: Computed by M. Longnecker using the R function qchisq$(1-\alpha, \text{df})$.

For level α two-tailed tests and $100(1-\alpha)\%$ C.I.s use value in columns headed by the numbers obtained by computing $1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$.

TABLE 7
(*continued*)

Right-Tail Probability (α)						
.10	**.05**	**.025**	**.01**	**.005**	**.001**	**df**
2.706	3.841	5.024	6.635	7.879	10.83	1
4.605	5.991	7.378	9.210	10.60	13.82	2
6.251	7.815	9.348	11.34	12.84	16.27	3
7.779	9.488	11.14	13.28	14.86	18.47	4
9.236	11.07	12.83	15.09	16.75	20.52	5
10.64	12.59	14.45	16.81	18.55	22.46	6
12.02	14.07	16.01	18.48	20.28	24.32	7
13.36	15.51	17.53	20.09	21.95	26.12	8
14.68	16.92	19.02	21.67	23.59	27.88	9
15.99	18.31	20.48	23.21	25.19	29.59	10
17.28	19.68	21.92	24.72	26.76	31.26	11
18.55	21.03	23.34	26.22	28.30	32.91	12
19.81	22.36	24.74	27.69	29.82	34.53	13
21.06	23.68	26.12	29.14	31.32	36.12	14
22.31	25.00	27.49	30.58	32.80	37.70	15
23.54	26.30	28.85	32.00	34.27	39.25	16
24.77	27.59	30.19	33.41	35.72	40.79	17
25.99	28.87	31.53	34.81	37.16	42.31	18
27.20	30.14	32.85	36.19	38.58	43.82	19
28.41	31.41	34.17	37.57	40.00	45.31	20
29.62	32.67	35.48	38.93	41.40	46.80	21
30.81	33.92	36.78	40.29	42.80	48.27	22
32.01	35.17	38.08	41.64	44.18	49.73	23
33.20	36.42	39.36	42.98	45.56	51.18	24
34.38	37.65	40.65	44.31	46.93	52.62	25
35.56	38.89	41.92	45.64	48.29	54.05	26
36.74	40.11	43.19	46.96	49.64	55.48	27
37.92	41.34	44.46	48.28	50.99	56.89	28
39.09	42.56	45.72	49.59	52.34	58.30	29
40.26	43.77	46.98	50.89	53.67	59.70	30
51.81	55.76	59.34	63.69	66.77	73.40	40
63.17	67.50	71.42	76.15	79.49	86.66	50
74.40	79.08	83.30	88.38	91.95	99.61	60
85.53	90.53	95.02	100.43	104.21	112.32	70
96.58	101.88	106.63	112.33	116.32	124.84	80
107.57	113.15	118.14	124.12	128.30	137.21	90
118.50	124.34	129.56	135.81	140.17	149.45	100
140.23	146.57	152.21	158.95	163.65	173.62	120
268.47	277.14	284.80	293.89	300.18	313.44	240

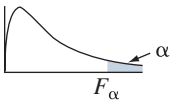


TABLE 8

Percentage points of the F distribution (df_2 between 1 and 6)

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
1	.25	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32
	.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
	.05	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
	.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6
	.01	4052.2	4999.5	5403.3	5624.6	5763.7	5859.0	5928.4	5981.0	6022.5	6055.8
2	.25	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38
	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
	.005	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4
	.001	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.4	999.4
3	.25	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44
	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
	.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
	.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69
	.001	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9	129.2
4	.25	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08
	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
	.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05
5	.25	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89
	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92
6	.25	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77
	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
	.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41

Source: Computed by M. Longnecker using the R function qf$(1 - \alpha, df_1, df_2)$.
Additional values can be obtained using the same R function.

TABLE 8

Percentage points of the F distribution (df_2 between 1 and 6)

df_1											
12	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**240**	**inf.**	α	df_2
9.41	9.49	9.58	9.63	9.67	9.71	9.76	9.80	9.83	9.85	.25	**1**
60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.19	63.33	.10	
243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	253.8	254.3	.05	
976.7	984.9	993.1	997.2	1001.4	1005.6	1009.8	1014.0	1016.1	1018.3	.025	
6106.3	6157.3	6208.7	6234.6	6260.6	6286.8	6313.0	6339.4	6352.6	6365.9	.01	
3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.47	3.48	.25	**2**
9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49	9.49	.10	
19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.49	19.50	.05	
39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.49	39.50	.025	
99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50	99.50	.01	
199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5	199.5	.005	
999.4	999.4	999.4	999.5	999.5	999.5	999.5	999.5	999.5	999.5	.001	
2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47	2.47	.25	**3**
5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.14	5.13	.10	
8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.54	8.53	.05	
14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.92	13.90	.025	
27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.17	26.13	.01	
43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.91	41.83	.005	
128.3	127.4	126.4	125.9	125.4	125.0	124.5	124.0	123.7	123.5	.001	
2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	.25	**4**
3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.77	3.76	.10	
5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.64	5.63	.05	
8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.28	8.26	.025	
14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.51	13.46	.01	
20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.40	19.32	.005	
47.41	46.76	46.10	45.77	45.43	45.09	44.75	44.40	44.23	44.05	.001	
1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87	1.87	.25	**5**
3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.11	3.10	.10	
4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.38	4.36	.05	
6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.04	6.02	.025	
9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.07	9.02	.01	
13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.21	12.14	.005	
26.42	25.91	25.39	25.13	24.87	24.60	24.33	24.06	23.92	23.79	.001	
1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74	1.74	.25	**6**
2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.73	2.72	.10	
4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.69	3.67	.05	
5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.88	4.85	.025	
7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.92	6.88	.01	
10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.94	8.88	.005	
17.99	17.56	17.12	16.90	16.67	16.44	16.21	15.98	15.86	15.75	.001	

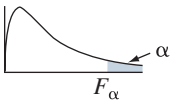


TABLE 8

Percentage points of the F distribution (df_2 between 7 and 12)

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
7	.25	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69
	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
	.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08
8	.25	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63
	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54
9	.25	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59
	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89
10	.25	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55
	.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75
11	.25	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52
	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
	.005	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92
12	.25	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50
	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29

TABLE 8

Percentage points of the F distribution (df_2 between 7 and 12)

df_1											
12	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**240**	**inf.**	α	df_2
1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65	1.65	.25	**7**
2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.48	2.47	.10	
3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.25	3.23	.05	
4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.17	4.14	.025	
6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.69	5.65	.01	
8.18	7.97	7.75	7.64	7.53	7.42	7.31	7.19	7.13	7.08	.005	
13.71	13.32	12.93	12.73	12.53	12.33	12.12	11.91	11.80	11.70	.001	
1.62	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58	1.58	.25	**8**
2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.30	2.29	.10	
3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.95	2.93	.05	
4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.70	3.67	.025	
5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.90	4.86	.01	
7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	6.01	5.95	.005	
11.19	10.84	10.48	10.30	10.11	9.92	9.73	9.53	9.43	9.33	.001	
1.58	1.57	1.56	1.56	1.55	1.54	1.54	1.53	1.53	1.53	.25	**9**
2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.17	2.16	.10	
3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.73	2.71	.05	
3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.36	3.33	.025	
5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.35	4.31	.01	
6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.24	5.19	.005	
9.57	9.24	8.90	8.72	8.55	8.37	8.19	8.00	7.91	7.81	.001	
1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.49	1.48	.25	**10**
2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.07	2.06	.10	
2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.56	2.54	.05	
3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.11	3.08	.025	
4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.95	3.91	.01	
5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.69	4.64	.005	
8.45	8.13	7.80	7.64	7.47	7.30	7.12	6.94	6.85	6.76	.001	
1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.45	1.45	.25	**11**
2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.99	1.97	.10	
2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.43	2.40	.05	
3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.91	2.88	.025	
4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.65	3.60	.01	
5.24	5.05	4.86	4.76	4.65	4.55	4.45	4.34	4.28	4.23	.005	
7.63	7.32	7.01	6.85	6.68	6.52	6.35	6.18	6.09	6.00	.001	
1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.43	1.43	1.42	.25	**12**
2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.92	1.90	.10	
2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.32	2.30	.05	
3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.76	2.72	.025	
4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.41	3.36	.01	
4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.96	3.90	.005	
7.00	6.71	6.40	6.25	6.09	5.93	5.76	5.59	5.51	5.42	.001	

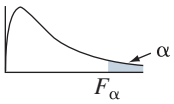


TABLE 8

Percentage points of the F distribution (df_2 between 13 and 18)

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
13	.25	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48
	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
	.005	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80
14	.25	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46
	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
	.005	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40
15	.25	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45
	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08
16	.25	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44
	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
	.005	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81
17	.25	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43
	.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
	.005	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58
18	.25	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42
	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
	.005	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39

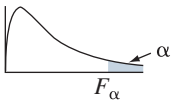
TABLE 8

Percentage points of the F distribution (df_2 between 13 and 18)

df_1											
12	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**240**	**inf.**	α	df_2
1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40	1.40	.25	**13**
2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.86	1.85	.10	
2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.23	2.21	.05	
3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.63	2.60	.025	
3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.21	3.17	.01	
4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.70	3.65	.005	
6.52	6.23	5.93	5.78	5.63	5.47	5.30	5.14	5.05	4.97	.001	
1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.39	1.38	1.38	.25	**14**
2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.81	1.80	.10	
2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.15	2.13	.05	
3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.52	2.49	.025	
3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.05	3.00	.01	
4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.49	3.44	.005	
6.13	5.85	5.56	5.41	5.25	5.10	4.94	4.77	4.69	4.60	.001	
1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.37	1.36	1.36	.25	**15**
2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.77	1.76	.10	
2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.09	2.07	.05	
2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.43	2.40	.025	
3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.91	2.87	.01	
4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.32	3.26	.005	
5.81	5.54	5.25	5.10	4.95	4.80	4.64	4.47	4.39	4.31	.001	
1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.35	1.34	.25	**16**
1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.73	1.72	.10	
2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.03	2.01	.05	
2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.35	2.32	.025	
3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.80	2.75	.01	
4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.17	3.11	.005	
5.55	5.27	4.99	4.85	4.70	4.54	4.39	4.23	4.14	4.06	.001	
1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.33	.25	**17**
1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.70	1.69	.10	
2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.99	1.96	.05	
2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.28	2.25	.025	
3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.70	2.65	.01	
3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	3.04	2.98	.005	
5.32	5.05	4.78	4.63	4.48	4.33	4.18	4.02	3.93	3.85	.001	
1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.32	.25	**18**
1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.67	1.66	.10	
2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.94	1.92	.05	
2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.22	2.19	.025	
3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.61	2.57	.01	
3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.93	2.87	.005	
5.13	4.87	4.59	4.45	4.30	4.15	4.00	3.84	3.75	3.67	.001	

TABLE 8

Percentage points of the F distribution (df_2 between 19 and 24)



		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
19	.25	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41
	.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
	.005	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22
20	.25	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40
	.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08
21	.25	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39
	.10	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
	.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
	.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
	.005	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95
22	.25	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39
	.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
	.005	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83
23	.25	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38
	.10	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
	.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
	.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
	.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73
24	.25	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38
	.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
	.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64

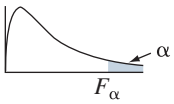
TABLE 8

Percentage points of the F distribution (df_2 between 19 and 24)

df_1											
12	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**240**	**inf.**	α	df_2
1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.30	.25	**19**
1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.65	1.63	.10	
2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.90	1.88	.05	
2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.17	2.13	.025	
3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.54	2.49	.01	
3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.83	2.78	.005	
4.97	4.70	4.43	4.29	4.14	3.99	3.84	3.68	3.60	3.51	.001	
1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.30	1.29	.25	**20**
1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.63	1.61	.10	
2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.87	1.84	.05	
2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.12	2.09	.025	
3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.47	2.42	.01	
3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.75	2.69	.005	
4.82	4.56	4.29	4.15	4.00	3.86	3.70	3.54	3.46	3.38	.001	
1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.29	1.28	.25	**21**
1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.60	1.59	.10	
2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.84	1.81	.05	
2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.08	2.04	.025	
3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.41	2.36	.01	
3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.67	2.61	.005	
4.70	4.44	4.17	4.03	3.88	3.74	3.58	3.42	3.34	3.26	.001	
1.37	1.36	1.34	1.33	1.32	1.31	1.30	1.29	1.28	1.28	.25	**22**
1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.59	1.57	.10	
2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.81	1.78	.05	
2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.04	2.00	.025	
3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.35	2.31	.01	
3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.60	2.55	.005	
4.58	4.33	4.06	3.92	3.78	3.63	3.48	3.32	3.23	3.15	.001	
1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28	1.28	1.27	.25	**23**
1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.57	1.55	.10	
2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.79	1.76	.05	
2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	2.01	1.97	.025	
3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.31	2.26	.01	
3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.54	2.48	.005	
4.48	4.23	3.96	3.82	3.68	3.53	3.38	3.22	3.14	3.05	.001	
1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28	1.27	1.26	.25	**24**
1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.55	1.53	.10	
2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.76	1.73	.05	
2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.97	1.94	.025	
3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.26	2.21	.01	
3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.49	2.43	.005	
4.39	4.14	3.87	3.74	3.59	3.45	3.29	3.14	3.05	2.97	.001	

TABLE 8

Percentage points of the F distribution (df_2 between 25 and 30)



		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
25	.25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37
	.10	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
	.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
	.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
	.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56
26	.25	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37
	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
	.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
	.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48
27	.25	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36
	.10	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
	.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
	.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
	.005	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41
28	.25	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36
	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
	.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
	.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41
	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35
29	.25	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35
	.10	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
	.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53
	.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
	.005	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38
	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29
30	.25	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35
	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34
	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24

TABLE 8

Percentage points of the F distribution (df_2 between 25 and 30)

df_1											
12	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**240**	**inf.**	α	df_2
1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.27	1.26	1.25	.25	**25**
1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.54	1.52	.10	
2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.74	1.71	.05	
2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.94	1.91	.025	
2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.22	2.17	.01	
3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.44	2.38	.005	
4.31	4.06	3.79	3.66	3.52	3.37	3.22	3.06	2.98	2.89	.001	
1.35	1.34	1.32	1.31	1.30	1.29	1.28	1.26	1.26	1.25	.25	**26**
1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.52	1.50	.10	
2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.72	1.69	.05	
2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.92	1.88	.025	
2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.18	2.13	.01	
3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.39	2.33	.005	
4.24	3.99	3.72	3.59	3.44	3.30	3.15	2.99	2.90	2.82	.001	
1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.25	1.24	.25	**27**
1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.51	1.49	.10	
2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.70	1.67	.05	
2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.89	1.85	.025	
2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.15	2.10	.01	
3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.35	2.29	.005	
4.17	3.92	3.66	3.52	3.38	3.23	3.08	2.92	2.84	2.75	.001	
1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24	1.24	.25	**28**
1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.50	1.48	.10	
2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.68	1.65	.05	
2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.87	1.83	.025	
2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.12	2.06	.01	
3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.31	2.25	.005	
4.11	3.86	3.60	3.46	3.32	3.18	3.02	2.86	2.78	2.69	.001	
1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.24	1.23	.25	**29**
1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.49	1.47	.10	
2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.67	1.64	.05	
2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.85	1.81	.025	
2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.09	2.03	.01	
3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.27	2.21	.005	
4.05	3.80	3.54	3.41	3.27	3.12	2.97	2.81	2.73	2.64	.001	
1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23	1.23	.25	**30**
1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.48	1.46	.10	
2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.65	1.62	.05	
2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.83	1.79	.025	
2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.06	2.01	.01	
3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.24	2.18	.005	
4.00	3.75	3.49	3.36	3.22	3.07	2.92	2.76	2.68	2.59	.001	

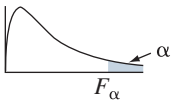


TABLE 8

Percentage points of the F distribution (df_2 at least 40)

		df_1									
df_2	α	1	2	3	4	5	6	7	8	9	10
40	.25	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33
	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
	.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87
60	.25	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30
	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54
90	.25	1.34	1.41	1.39	1.37	1.35	1.33	1.32	1.31	1.30	1.29
	.10	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
	.05	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	.025	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19
	.01	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
	.005	8.28	5.62	4.57	3.99	3.62	3.35	3.15	3.00	2.87	2.77
	.001	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34
120	.25	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
	.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
	.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
	.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71
	.001	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24
240	.25	1.33	1.39	1.38	1.36	1.34	1.32	1.30	1.29	1.27	1.27
	.10	2.73	2.32	2.10	1.97	1.87	1.80	1.74	1.70	1.65	1.63
	.05	3.88	3.03	2.64	2.41	2.25	2.14	2.04	1.98	1.92	1.87
	.025	5.09	3.75	3.17	2.84	2.62	2.46	2.34	2.25	2.17	2.10
	.01	6.74	4.69	3.86	3.40	3.09	2.88	2.71	2.59	2.48	2.40
	.005	8.03	5.42	4.38	3.82	3.45	3.19	2.99	2.84	2.71	2.61
	.001	11.10	7.11	5.60	4.78	4.25	3.89	3.62	3.41	3.24	3.09
inf.	.25	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25
	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32
	.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52
	.001	10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96

TABLE 8

Percentage points of the F distribution (df_2 at least 40)

df_1											
12	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**240**	**inf.**	α	df_2
1.31	1.30	1.28	1.26	1.25	1.24	1.22	1.21	1.20	1.19	.25	**40**
1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.40	1.38	.10	
2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.54	1.51	.05	
2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.68	1.64	.025	
2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.86	1.80	.01	
2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	2.00	1.93	.005	
3.64	3.40	3.14	3.01	2.87	2.73	2.57	2.41	2.32	2.23	.001	
1.29	1.27	1.25	1.24	1.22	1.21	1.19	1.17	1.16	1.15	.25	**60**
1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.32	1.29	.10	
1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.43	1.39	.05	
2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.53	1.48	.025	
2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.67	1.60	.01	
2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.76	1.69	.005	
3.32	3.08	2.83	2.69	2.55	2.41	2.25	2.08	1.99	1.89	.001	
1.27	1.25	1.23	1.22	1.20	1.19	1.17	1.15	1.13	1.12	.25	**90**
1.62	1.56	1.50	1.47	1.43	1.39	1.35	1.29	1.26	1.23	.10	
1.86	1.78	1.69	1.64	1.59	1.53	1.46	1.39	1.35	1.30	.05	
2.09	1.98	1.86	1.80	1.73	1.66	1.58	1.48	1.43	1.37	.025	
2.39	2.24	2.09	2.00	1.92	1.82	1.72	1.60	1.53	1.46	.01	
2.61	2.44	2.25	2.15	2.05	1.94	1.82	1.68	1.61	1.52	.005	
3.11	2.88	2.63	2.50	2.36	2.21	2.05	1.87	1.77	1.66	.001	
1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.12	1.10	.25	**120**
1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.23	1.19	.10	
1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.31	1.25	.05	
2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.38	1.31	.025	
2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.46	1.38	.01	
2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.52	1.43	.005	
3.02	2.78	2.53	2.40	2.26	2.11	1.95	1.77	1.66	1.54	.001	
1.25	1.23	1.21	1.19	1.18	1.16	1.14	1.11	1.09	1.07	.25	**240**
1.57	1.52	1.45	1.42	1.38	1.33	1.28	1.22	1.18	1.13	.10	
1.79	1.71	1.61	1.56	1.51	1.44	1.37	1.29	1.24	1.17	.05	
2.00	1.89	1.77	1.70	1.63	1.55	1.46	1.35	1.29	1.21	.025	
2.26	2.11	1.96	1.87	1.78	1.68	1.57	1.43	1.35	1.25	.01	
2.45	2.28	2.09	1.99	1.89	1.77	1.64	1.49	1.40	1.28	.005	
2.88	2.65	2.40	2.26	2.12	1.97	1.80	1.61	1.49	1.35	.001	
1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.06	1.00	.25	**inf.**
1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.12	1.00	.10	
1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.15	1.00	.05	
1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.19	1.00	.025	
2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.22	1.00	.01	
2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.25	1.00	.005	
2.74	2.51	2.27	2.13	1.99	1.84	1.66	1.45	1.31	1.00	.001	

TABLE 9
Values of $y = 2 \arcsin \sqrt{\hat{\pi}}$

$\hat{\pi}$	y	$\hat{\pi}$	y	$\hat{\pi}$	y	$\hat{\pi}$	y	$\hat{\pi}$	y
.001	.0633	.041	.4078	.36	1.2870	.76	2.1177	.971	2.7993
.002	.0895	.042	.4128	.37	1.3078	.77	2.1412	.972	2.8053
.003	.1096	.043	.4178	.38	1.3284	.78	2.1652	.973	2.8115
.004	.1266	.044	.4227	.39	1.3490	.79	2.1895	.974	2.8177
.005	.1415	.045	.4275	.40	1.3694	.80	2.2143	.975	2.8240
.006	.1551	.046	.4323	.41	1.3898	.81	2.2395	.976	2.8305
.007	.1675	.047	.4371	.42	1.4101	.82	2.2653	.977	2.8371
.008	.1791	.048	.4418	.43	1.4303	.83	2.2916	.978	2.8438
.009	.1900	.049	.4464	.44	1.4505	.84	2.3186	.979	2.8507
.010	.2003	.050	.4510	.45	1.4706	.85	2.3462	.980	2.8578
.011	.2101	.06	.4949	.46	1.4907	.86	2.3746	.981	2.8650
.012	.2195	.07	.5355	.47	1.5108	.87	2.4039	.982	2.8725
.013	.2285	.08	.5735	.48	1.5308	.88	2.4341	.983	2.8801
.014	.2372	.09	.6094	.49	1.5508	.89	2.4655	.984	2.8879
.015	.2456	.10	.6435	.50	1.5708	.90	2.4981	.985	2.8960
.016	.2537	.11	.6761	.51	1.5908	.91	2.5322	.986	2.9044
.017	.2615	.12	.7075	.52	1.6108	.92	2.5681	.987	2.9131
.018	.2691	.13	.7377	.53	1.6308	.93	2.6061	.988	2.9221
.019	.2766	.14	.7670	.54	1.6509	.94	2.6467	.989	2.9314
.020	.2838	.15	.7954	.55	1.6710	.95	2.6906	.990	2.9413
.021	.2909	.16	.8230	.56	1.6911	.951	2.6952	.991	2.9516
.022	.2978	.17	.8500	.57	1.7113	.952	2.6998	.992	2.9625
.023	.3045	.18	.8763	.58	1.7315	.953	2.7045	.993	2.9741
.024	.3111	.19	.9021	.59	1.7518	.954	2.7093	.994	2.9865
.025	.3176	.20	.9273	.60	1.7722	.955	2.7141	.995	3.0001
.026	.3239	.21	.9521	.61	1.7926	.956	2.7189	.996	3.0150
.027	.3301	.22	.9764	.62	1.8132	.957	2.7238	.997	3.0320
.028	.3363	.23	1.0004	.63	1.8338	.958	2.7288	.998	3.0521
.029	.3423	.24	1.0239	.64	1.8546	.959	2.7338	.999	3.0783
.030	.3482	.25	1.0472	.65	1.8755	.960	2.7389		
.031	.3540	.26	1.0701	.66	1.8965	.961	2.7440		
.032	.3597	.27	1.0928	.67	1.9177	.962	2.7492		
.033	.3653	.28	1.1152	.68	1.9391	.963	2.7545		
.034	.3709	.29	1.1374	.69	1.9606	.964	2.7598		
.035	.3764	.30	1.1593	.70	1.9823	.965	2.7652		
.036	.3818	.31	1.1810	.71	2.0042	.966	2.7707		
.037	.3871	.32	1.2025	.72	2.0264	.967	2.7762		
.038	.3924	.33	1.2239	.73	2.0488	.968	2.7819		
.039	.3976	.34	1.2451	.74	2.0715	.969	2.7876		
.040	.4027	.35	1.2661	.75	2.0944	.970	2.7934		

Source: Computed by M. Longnecker using the R function 2*asin(sqrt ($\hat{\pi}$)).

TABLE 10

Percentage points of the Studentized range

Error df	α	t = Number of Treatment Means									
		2	3	4	5	6	7	8	9	10	11
5	.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
	.01	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48
6	.05	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65
	.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30
7	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30
	.01	4.95	5.92	6.54	7.00	7.37	7.68	7.94	8.17	8.37	8.55
8	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
	.01	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03
9	.05	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87
	.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	7.65
10	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72
	.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36
11	.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61
	.01	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13
12	.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51
	.01	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94
13	.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43
	.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79
14	.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
	.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66
15	.05	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31
	.01	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55
16	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26
	.01	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46
17	.05	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21
	.01	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38
18	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17
	.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31
19	.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14
	.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25
20	.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11
	.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19
24	.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01
	.01	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02
30	.05	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92
	.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85
40	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82
	.01	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.69
60	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73
	.01	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.53
120	.05	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64
	.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.37
∞	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55
	.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23

Source: Computed by M. Longnecker using the R function qtukey($1 - \alpha$, t, df).

TABLE 10
(*continued*)

Error df	t = Number of Treatment Means									
	12	13	14	15	16	17	18	19	20	α
5	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21	.05
	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93	.01
6	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59	.05
	9.48	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54	.01
7	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.10	7.17	.05
	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65	.01
8	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87	.05
	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03	.01
9	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64	.05
	7.78	7.91	8.03	8.13	8.23	8.33	8.41	8.49	8.57	.01
10	5.83	5.93	6.03	6.11	6.19	6.27	6.34	6.40	6.47	.05
	7.49	7.60	7.71	7.81	7.91	7.99	8.08	8.15	8.23	.01
11	5.71	5.81	5.90	5.98	6.06	6.13	6.20	6.27	6.33	.05
	7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95	.01
12	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21	.05
	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73	.01
13	5.53	5.63	5.71	5.79	5.86	5.93	5.99	6.05	6.11	.05
	6.90	7.01	7.10	7.19	7.27	7.35	7.42	7.48	7.55	.01
14	5.46	5.55	5.64	5.71	5.79	5.85	5.91	5.97	6.03	.05
	6.77	6.87	6.96	7.05	7.13	7.20	7.27	7.33	7.39	.01
15	5.40	5.49	5.57	5.65	5.72	5.78	5.85	5.90	5.96	.05
	6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26	.01
16	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90	.05
	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15	.01
17	5.31	5.39	5.47	5.54	5.61	5.67	5.73	5.79	5.84	.05
	6.48	6.57	6.66	6.73	6.81	6.87	6.94	7.00	7.05	.01
18	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79	.05
	6.41	6.50	6.58	6.65	6.73	6.79	6.85	6.91	6.97	.01
19	5.23	5.31	5.39	5.46	5.53	5.59	5.65	5.70	5.75	.05
	6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89	.01
20	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71	.05
	6.28	6.37	6.45	6.52	6.59	6.65	6.71	6.77	6.82	.01
24	5.10	5.18	5.25	5.32	5.38	5.44	5.49	5.55	5.59	.05
	6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61	.01
30	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.47	.05
	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41	.01
40	4.90	4.98	5.04	5.11	5.16	5.22	5.27	5.31	5.36	.05
	5.76	5.83	5.90	5.96	6.02	6.07	6.12	6.16	6.21	.01
60	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24	.05
	5.60	5.67	5.73	5.78	5.84	5.89	5.93	5.97	6.01	.01
120	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13	.05
	5.44	5.50	5.56	5.61	5.66	5.71	5.75	5.79	5.83	.01
∞	4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01	.05
	5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.61	5.65	.01

TABLE 11

Percentage points for Dunnett's test: $d_\alpha(k, \nu)$

	$\alpha = .05$ (one-sided)												
ν	$k = 2$	3	4	5	6	7	8	9	10	11	12	15	20
5	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30	3.36	3.41	3.45	3.57	3.72
6	2.34	2.56	2.71	2.83	2.92	3.00	3.07	3.12	3.17	3.22	3.26	3.37	3.50
7	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.01	3.05	3.10	3.13	3.23	3.36
8	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92	2.96	3.01	3.04	3.14	3.25
9	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86	2.90	2.94	2.97	3.06	3.18
10	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81	2.85	2.89	2.92	3.01	3.12
11	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77	2.81	2.85	2.88	2.96	3.07
12	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74	2.78	2.81	2.84	2.93	3.03
13	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71	2.75	2.78	2.82	2.90	3.00
14	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69	2.72	2.76	2.79	2.87	2.97
15	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67	2.70	2.74	2.77	2.85	2.95
16	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65	2.69	2.72	2.75	2.83	2.93
17	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64	2.67	2.71	2.74	2.81	2.91
18	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62	2.66	2.69	2.72	2.80	2.89
19	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61	2.65	2.68	2.71	2.79	2.88
20	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60	2.64	2.67	2.70	2.77	2.87
24	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57	2.60	2.64	2.66	2.74	2.83
30	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54	2.57	2.60	2.63	2.70	2.79
40	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51	2.54	2.57	2.60	2.67	2.75
60	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48	2.51	2.54	2.56	2.63	2.72
120	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45	2.48	2.51	2.53	2.60	2.68
∞	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42	2.45	2.48	2.50	2.56	2.64

From C. W. Dunnett. (1955). "A Multiple Comparison Procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association* 50, 1112–1118. Reprinted with permission from *Journal of the American Statistical Association*. Copyright 1955 by the American Statistical Association. All rights reserved. C. W. Dunnett. (1964). "New Tables for Multiple Comparisons with a Control," *Biometrics* 20, 482–491. Also additional tables produced by C. W. Dunnett in 1980.

TABLE 11
(*continued*)

$\alpha = .01$ (one-sided)													
ν	$k = 2$	3	4	5	6	7	8	9	10	11	12	15	20
5	3.90	4.21	4.43	4.60	4.73	4.85	4.94	5.03	5.11	5.17	5.24	5.39	5.59
6	3.61	3.88	4.07	4.21	4.33	4.43	4.51	4.59	4.64	4.70	4.76	4.89	5.06
7	3.42	3.66	3.83	3.96	4.07	4.15	4.23	4.30	4.35	4.40	4.45	4.57	4.72
8	3.29	3.51	3.67	3.79	3.88	3.96	4.03	4.09	4.14	4.19	4.23	4.34	4.48
9	3.19	3.40	3.55	3.66	3.75	3.82	3.89	3.94	3.99	4.04	4.08	4.18	4.31
10	3.11	3.31	3.45	3.56	3.64	3.71	3.78	3.83	3.88	3.92	3.96	4.06	4.18
11	3.06	3.25	3.38	3.48	3.56	3.63	3.69	3.74	3.79	3.83	3.86	3.96	4.08
12	3.01	3.19	3.32	3.42	3.50	3.56	3.62	3.67	3.71	3.75	3.79	3.88	3.99
13	2.97	3.15	3.27	3.37	3.44	3.51	3.56	3.61	3.65	3.69	3.73	3.81	3.92
14	2.94	3.11	3.23	3.32	3.40	3.46	3.51	3.56	3.60	3.64	3.67	3.76	3.87
15	2.91	3.08	3.20	3.29	3.36	3.42	3.47	3.52	3.56	3.60	3.63	3.71	3.82
16	2.88	3.05	3.17	3.26	3.33	3.39	3.44	3.48	3.52	3.56	3.59	3.67	3.78
17	2.86	3.03	3.14	3.23	3.30	3.36	3.41	3.45	3.49	3.53	3.56	3.64	3.74
18	2.84	3.01	3.12	3.21	3.27	3.33	3.38	3.42	3.46	3.50	3.53	3.61	3.71
19	2.83	2.99	3.10	3.18	3.25	3.31	3.36	3.40	3.44	3.47	3.50	3.58	3.68
20	2.81	2.97	3.08	3.17	3.23	3.29	3.34	3.38	3.42	3.45	3.48	3.56	3.65
24	2.77	2.92	3.03	3.11	3.17	3.22	3.27	3.31	3.35	3.38	3.41	3.48	3.57
30	2.72	2.87	2.97	3.05	3.11	3.16	3.21	3.24	3.28	3.31	3.34	3.41	3.50
40	2.68	2.82	2.92	2.99	3.05	3.10	3.14	3.18	3.21	3.24	3.27	3.34	3.42
60	2.64	2.78	2.87	2.94	3.00	3.04	3.08	3.12	3.15	3.18	3.20	3.27	3.35
120	2.60	2.73	2.82	2.89	2.94	2.99	3.03	3.06	3.09	3.12	3.14	3.20	3.28
∞	2.56	2.68	2.77	2.84	2.89	2.93	2.97	3.00	3.03	3.06	3.08	3.14	3.21

TABLE 11
(*continued*)

$\alpha = .05$ (two-sided)													
ν	$k = 2$	3	4	5	6	7	8	9	10	11	12	15	20
5	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97	4.03	4.09	4.14	4.26	4.42
6	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71	3.76	3.81	3.86	3.97	4.11
7	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53	3.58	3.63	3.67	3.78	3.91
8	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41	3.46	3.50	3.54	3.64	3.76
9	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32	3.36	3.40	3.44	3.53	3.65
10	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24	3.29	3.33	3.36	3.45	3.57
11	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19	3.23	3.27	3.30	3.39	3.50
12	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14	3.18	3.22	3.25	3.34	3.45
13	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10	3.14	3.18	3.21	3.29	3.40
14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07	3.11	3.14	3.18	3.26	3.36
15	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04	3.08	3.12	3.15	3.23	3.33
16	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02	3.06	3.09	3.12	3.20	3.30
17	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00	3.03	3.07	3.10	3.18	3.27
18	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98	3.01	3.05	3.08	3.16	3.25
19	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96	3.00	3.03	3.06	3.14	3.23
20	2.38	2.54	2.65	2.73	2.80	2.86	2.90	2.95	2.98	3.02	3.05	3.12	3.22
24	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90	2.94	2.97	3.00	3.07	3.16
30	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86	2.89	2.92	2.95	3.02	3.11
40	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81	2.85	2.87	2.90	2.97	3.06
60	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77	2.80	2.83	2.86	2.92	3.00
120	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73	2.76	2.79	2.81	2.87	2.95
∞	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69	2.72	2.74	2.77	2.83	2.91

TABLE 11
(*continued*)

	$\alpha = .01$ (two-sided)												
ν	$k = 2$	3	4	5	6	7	8	9	10	11	12	15	20
5	4.63	4.98	5.22	5.41	5.56	5.69	5.80	5.89	5.98	6.05	6.12	6.30	6.52
6	4.21	4.51	4.71	4.87	5.00	5.10	5.20	5.28	5.35	5.41	5.47	5.62	5.81
7	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89	4.95	5.01	5.06	5.19	5.36
8	3.77	4.00	4.17	4.29	4.40	4.48	4.56	4.62	4.68	4.73	4.78	4.90	5.05
9	3.63	3.85	4.01	4.12	4.22	4.30	4.37	4.43	4.48	4.53	4.57	4.68	4.82
10	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28	4.33	4.37	4.42	4.52	4.65
11	3.45	3.65	3.79	3.89	3.98	4.05	4.11	4.16	4.21	4.25	4.29	4.39	4.52
12	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07	4.12	4.16	4.19	4.29	4.41
13	3.33	3.52	3.65	3.74	3.82	3.89	3.94	3.99	4.04	4.08	4.11	4.20	4.32
14	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93	3.97	4.01	4.05	4.13	4.24
15	3.25	3.43	3.55	3.64	3.71	3.78	3.83	3.88	3.92	3.95	3.99	4.07	4.18
16	3.22	3.39	3.51	3.60	3.67	3.73	3.78	3.83	3.87	3.91	3.94	4.02	4.13
17	3.19	3.36	3.47	3.56	3.63	3.69	3.74	3.79	3.83	3.86	3.90	3.98	4.08
18	3.17	3.33	3.44	3.53	3.60	3.66	3.71	3.75	3.79	3.83	3.86	3.94	4.04
19	3.15	3.31	3.42	3.50	3.57	3.63	3.68	3.72	3.76	3.79	3.83	3.90	4.00
20	3.13	3.29	3.40	3.48	3.55	3.60	3.65	3.69	3.73	3.77	3.80	3.87	3.97
24	3.07	3.22	3.32	3.40	3.47	3.52	3.57	3.61	3.64	3.68	3.70	3.78	3.87
30	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52	3.56	3.59	3.62	3.69	3.78
40	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44	3.48	3.51	3.53	3.60	3.68
60	2.90	3.03	3.12	3.19	3.25	3.29	3.33	3.37	3.40	3.42	3.45	3.51	3.59
120	2.85	2.97	3.06	3.12	3.18	3.22	3.26	3.29	3.32	3.35	3.37	3.43	3.51
∞	2.79	2.92	3.00	3.06	3.11	3.15	3.19	3.22	3.25	3.27	3.29	3.35	3.42

TABLE 12

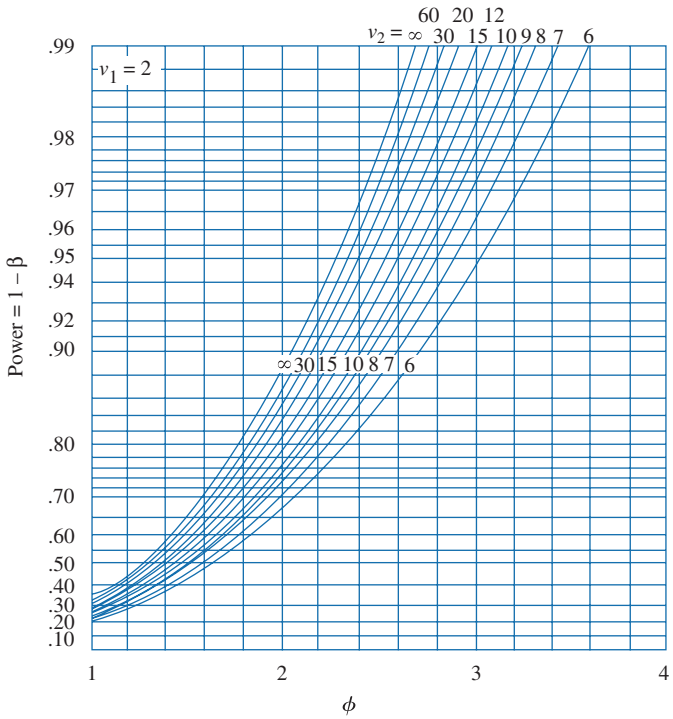
Random numbers

Line/ Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969	99570	91291	90700
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
6	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
7	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
8	96301	91977	05463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
9	89579	14342	63661	10281	17453	18103	57740	84378	25331	12566	58678	44947	05585	56941
10	85475	36857	53342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
11	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
12	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
13	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
14	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81056	97735	85977	29372	74461	28551	90707
16	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
17	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
18	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
19	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
21	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
22	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
23	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
24	29334	27001	87637	87308	58731	00256	45834	15398	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953

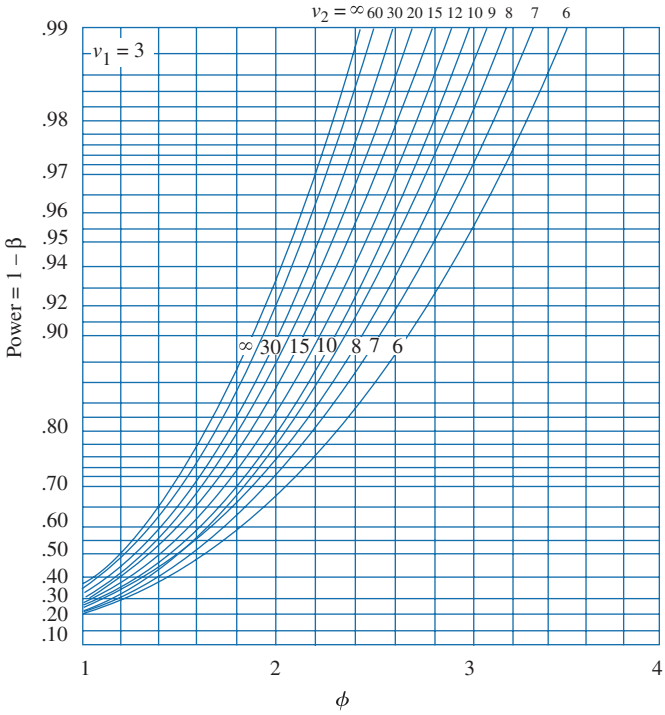
Abridged from William H. Beyer, ed., *Handbook of Tables for Probability and Statistics,* 2nd ed. © The Chemical Rubber Co., 1968.
Used by permission of CRC Press, Inc.

TABLE 13

F test power curves for AOV ($\alpha = .05$, $t = 3$)



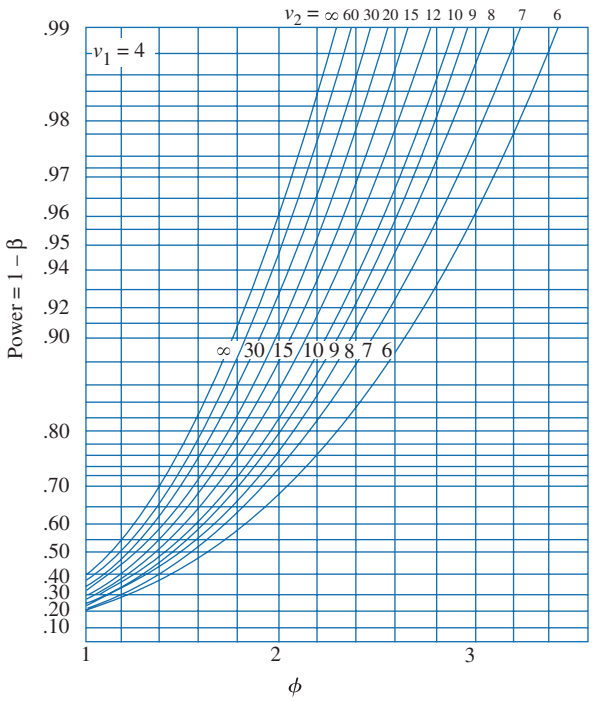
F test power curves for AOV ($\alpha = .05$, $t = 4$)



Data from *Biometrika Tables for Statisticians,* 1966, edited by E. S. Pearson and H. O. Hartley. Cambridge University, New York.

TABLE 13

F test power curves for AOV ($\alpha = .05, t = 5$)



F test power curves for AOV ($\alpha = .05, t = 6$)

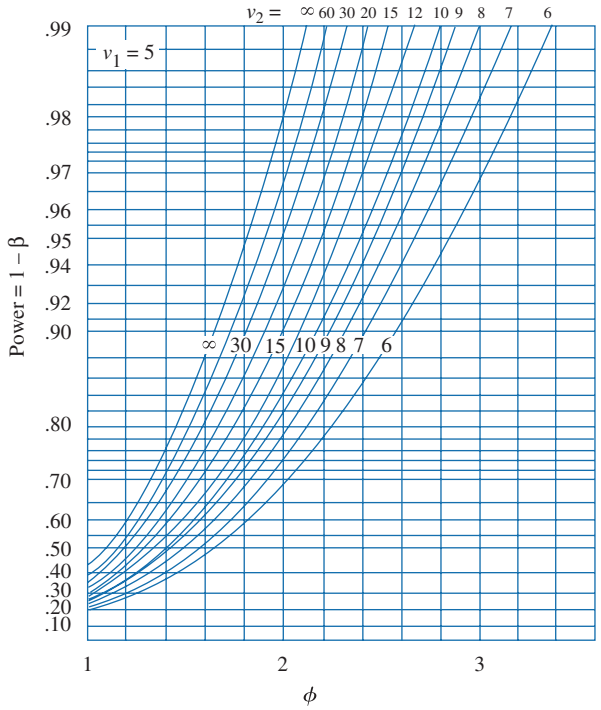
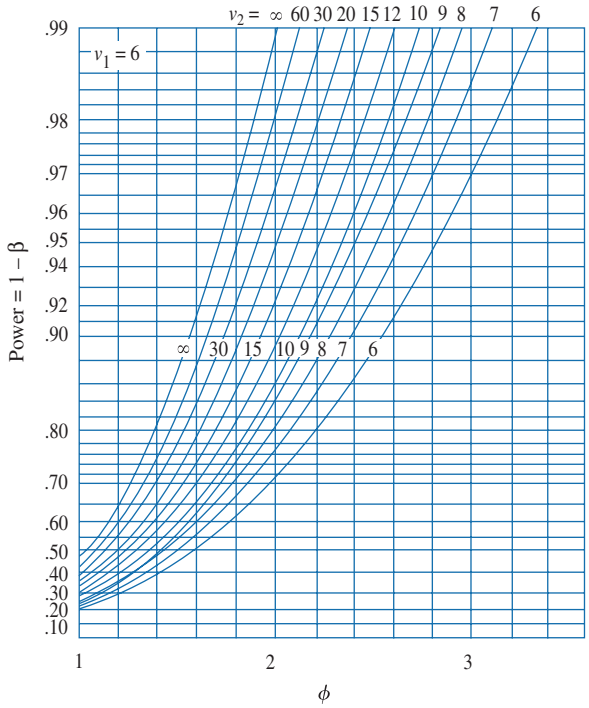


TABLE 13

F test power curves for AOV ($\alpha = .05, t = 7$)



F test power curves for AOV ($\alpha = .05, t = 8$)

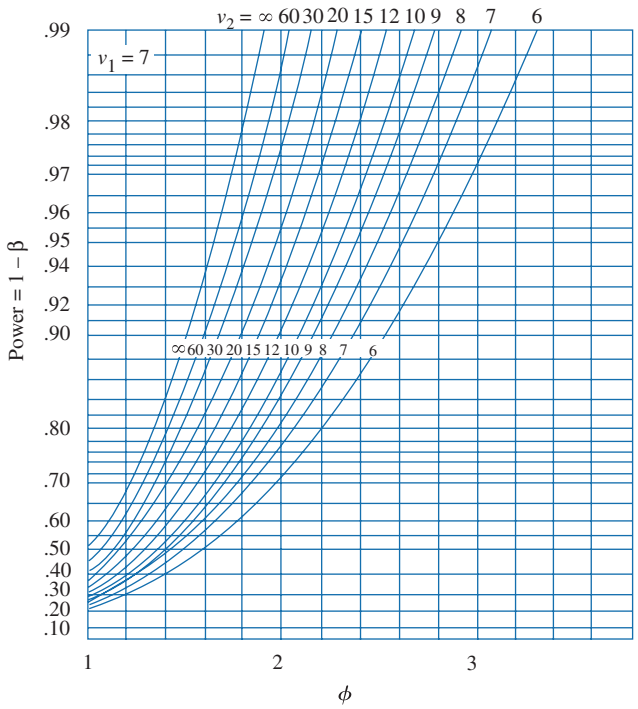


TABLE 13

F test power curves for AOV ($\alpha = .05$, $t = 9$)

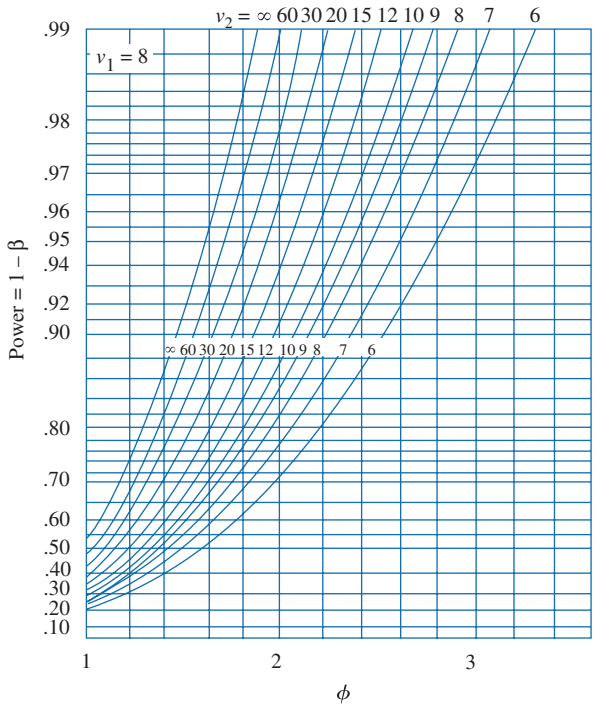


TABLE 14

Poisson probabilities $P(Y = y)$ (μ between .1 and 4.0)

					μ					
y	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005

					μ					
y	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009

					μ					
y	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1494
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680
5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027
10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008
11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002

					μ					
y	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1733	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954
5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042

Source: Computed by M. Longnecker using the R function dpois(y, μ).
Additional values can be obtained using the same R function.

TABLE 14

Poisson probabilities $P(Y = y)$ (μ between 3.1 and 10.0)

					μ					
y	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
7	.0246	.0278	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002

					μ					
y	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0281	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0013	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002

					μ					
y	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0
0	.0041	.0025	.0015	.0009	.0006	.0003	.0002	.0001	.0001	.0000
1	.0225	.0149	.0098	.0064	.0041	.0027	.0017	.0011	.0007	.0005
2	.0618	.0446	.0318	.0223	.0156	.0107	.0074	.0050	.0034	.0023
3	.1133	.0892	.0688	.0521	.0389	.0286	.0208	.0150	.0107	.0076
4	.1558	.1339	.1118	.0912	.0729	.0573	.0443	.0337	.0254	.0189
5	.1714	.1606	.1454	.1277	.1094	.0916	.0752	.0607	.0483	.0378
6	.1571	.1606	.1575	.1490	.1367	.1221	.1066	.0911	.0764	.0631
7	.1234	.1377	.1462	.1490	.1465	.1396	.1294	.1171	.1037	.0901
8	.0849	.1033	.1188	.1304	.1373	.1396	.1375	.1318	.1232	.1126
9	.0519	.0688	.0858	.1014	.1144	.1241	.1299	.1318	.1300	.1251
10	.0285	.0413	.0558	.0710	.0858	.0993	.1104	.1186	.1235	.1251
11	.0143	.0225	.0330	.0452	.0585	.0722	.0853	.0970	.1067	.1137
12	.0065	.0113	.0179	.0263	.0366	.0481	.0604	.0728	.0844	.0948
13	.0028	.0052	.0089	.0142	.0211	.0296	.0395	.0504	.0617	.0729
14	.0011	.0022	.0041	.0071	.0113	.0169	.0240	.0324	.0419	.0521
15	.0004	.0009	.0018	.0033	.0057	.0090	.0136	.0194	.0265	.0347

TABLE 14

Poisson probabilities $P(Y = y)$ (μ between 5.5 and 20.0)

					μ					
y	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0
16	.0001	.0003	.0007	.0014	.0026	.0045	.0072	.0109	.0157	.0217
17	.0000	.0001	.0003	.0006	.0012	.0021	.0036	.0058	.0088	.0128
18	.0000	.0000	.0001	.0002	.0005	.0009	.0017	.0029	.0046	.0071
19	.0000	.0000	.0000	.0001	.0002	.0004	.0008	.0014	.0023	.0037
20	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0006	.0011	.0019
21	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0003	.0005	.0009
22	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0004
23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002

					μ					
y	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0010	.0004	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000
3	.0037	.0018	.0008	.0004	.0002	.0001	.0000	.0000	.0000	.0000
4	.0102	.0053	.0027	.0013	.0006	.0003	.0001	.0001	.0000	.0000
5	.0224	.0127	.0070	.0037	.0019	.0010	.0005	.0002	.0001	.0001
6	.0411	.0255	.0152	.0087	.0048	.0026	.0014	.0007	.0004	.0002
7	.0646	.0437	.0281	.0174	.0104	.0060	.0034	.0019	.0010	.0005
8	.0888	.0655	.0457	.0304	.0194	.0120	.0072	.0042	.0024	.0013
9	.1085	.0874	.0661	.0473	.0324	.0213	.0135	.0083	.0050	.0029
10	.1194	.1048	.0859	.0663	.0486	.0341	.0230	.0150	.0095	.0058
11	.1194	.1144	.1015	.0844	.0663	.0496	.0355	.0245	.0164	.0106
12	.1094	.1144	.1099	.0984	.0829	.0661	.0504	.0368	.0259	.0176
13	.0926	.1056	.1099	.1060	.0956	.0814	.0658	.0509	.0378	.0271
14	.0728	.0905	.1021	.1060	.1024	.0930	.0800	.0655	.0514	.0387
15	.0534	.0724	.0885	.0989	.1024	.0992	.0906	.0786	.0650	.0516
16	.0367	.0543	.0719	.0866	.0960	.0992	.0963	.0884	.0772	.0646
17	.0237	.0383	.0550	.0713	.0847	.0934	.0963	.0936	.0863	.0760
18	.0145	.0255	.0397	.0554	.0706	.0830	.0909	.0936	.0911	.0844
19	.0084	.0161	.0272	.0409	.0557	.0699	.0814	.0887	.0911	.0888
20	.0046	.0097	.0177	.0286	.0418	.0559	.0692	.0798	.0866	.0888
21	.0024	.0055	.0109	.0191	.0299	.0426	.0560	.0684	.0783	.0846
22	.0012	.0030	.0065	.0121	.0204	.0310	.0433	.0560	.0676	.0769
23	.0006	.0016	.0037	.0074	.0133	.0216	.0320	.0438	.0559	.0669
24	.0003	.0008	.0020	.0043	.0083	.0144	.0226	.0328	.0442	.0557
25	.0001	.0004	.0010	.0024	.0050	.0092	.0154	.0237	.0336	.0446
26	.0000	.0002	.0005	.0013	.0029	.0057	.0101	.0164	.0246	.0343
27	.0000	.0001	.0002	.0007	.0016	.0034	.0063	.0109	.0173	.0254
28	.0000	.0000	.0001	.0003	.0009	.0019	.0038	.0070	.0117	.0181
29	.0000	.0000	.0001	.0002	.0004	.0011	.0023	.0044	.0077	.0125
30	.0000	.0000	.0000	.0001	.0002	.0006	.0013	.0026	.0049	.0083
31	.0000	.0000	.0000	.0000	.0001	.0003	.0007	.0015	.0030	.0054
32	.0000	.0000	.0000	.0000	.0001	.0001	.0004	.0009	.0018	.0034
33	.0000	.0000	.0000	.0000	0000	.0001	.0002	.0005	.0010	.0020

TABLE 15

Percentage points of the normal probability plot correlation coefficent, r

n/α =	.005	.01	.025	.05	.10	.25	.50	.75	.90	.95	.975	.99	.995
10	.860	.876	.900	.917	.934	.954	.970	.981	.987	.990	.992	.994	.995
11	.868	.883	.906	.922	.938	.957	.972	.982	.988	.990	.992	.994	.995
12	.875	.889	.912	.926	.941	.959	.973	.982	.988	.990	.992	.994	.995
13	.882	.895	.917	.931	.944	.962	.975	.983	.988	.991	.993	.994	.995
14	.888	.901	.921	.934	.947	.964	.976	.984	.989	.991	.993	.994	.995
15	.894	.907	.925	.937	.950	.965	.977	.984	.989	.991	.993	.994	.995
16	.889	.912	.928	.940	.952	.967	.978	.985	.989	.991	.993	.994	.995
17	.903	.916	.931	.942	.954	.968	.979	.986	.990	.992	.993	.994	.995
18	.907	.919	.934	.945	.956	.969	.979	.986	.990	.992	.993	.995	.995
19	.909	.923	.937	.947	.958	.971	.980	.987	.990	.992	.993	.995	.995
20	.912	.925	.939	.950	.960	.972	.981	.987	.991	.992	.994	.995	.995
21	.914	.928	.942	.952	.961	.973	.981	.987	.991	.993	.994	.995	.996
22	.918	.930	.944	.954	.962	.974	.982	.988	.991	.993	.994	.995	.996
23	.922	.933	.947	.955	.964	.975	.983	.988	.991	.993	.994	.995	.996
24	.926	.936	.949	.957	.965	.975	.983	.988	.992	.993	.994	.995	.996
25	.928	.937	.950	.958	.966	.976	.984	.989	.992	.993	.994	.995	.996
26	.930	.939	.952	.959	.967	.977	.984	.989	.992	.993	.994	.995	.996
27	.932	.941	.953	.960	.968	.977	.984	.989	.992	.994	.995	.995	.996
28	.934	.943	.955	.962	.969	.978	.985	.990	.992	.994	.995	.995	.996
29	.937	.945	.956	.962	.969	.979	.985	.990	.992	.994	.995	.995	.996
30	.938	.947	.957	.964	.970	.979	.986	.990	.993	.994	.995	.996	.996
35	.943	.952	.961	.968	.974	.982	.987	.991	.993	.995	.995	.996	.997
40	.949	.958	.966	.972	.977	.983	.988	.992	.994	.995	.996	.996	.997
45	.955	.961	.969	.974	.978	.985	.989	.993	.994	.995	.996	.997	.997
50	.959	.965	.972	.977	.981	.986	.990	.993	.995	.996	.996	.997	.997
55	.962	.967	.974	.978	.982	.987	.991	.994	.995	.996	.997	.997	.997
60	.965	.970	.976	.980	.983	.988	.991	.994	.995	.996	.997	.997	.998
65	.967	.972	.977	.981	.984	.989	.992	.994	.996	.996	.997	.997	.998
70	.969	.974	.978	.982	.985	.989	.993	.995	.996	.997	.997	.998	.998
75	.971	.975	.979	.983	.986	.990	.993	.995	.996	.997	.997	.998	.998
80	.973	.976	.980	.984	.987	.991	.993	.995	.996	.997	.997	.998	.998
85	.974	.977	.981	.985	.987	.991	.994	.995	.997	.997	.997	.998	.998
90	.976	.978	.982	.985	.988	.991	.994	.996	.997	.997	.998	.998	.998
95	.977	.979	.983	.986	.989	.992	.994	.996	.997	.997	.998	.998	.998
100	.979	.981	.984	.987	.989	.992	.994	.996	.997	.998	.998	.998	.998

From J. J. Filliben. (1975). "The Probability Plot Correlation Coefficient Test for Normality," *Technometrics* 17, 111–117.

Answers to Selected Exercises*

Chapter 1: Statistics and the Scientific Method

1.6 **a.** All freshman at the university
b. Just the freshman enrolled in HIST 101
c. The students taking HIST 101 may have a different level of knowledge about history than the general student body.
d. The initial lectures would certainly give the students information about the original 13 colonies and hence make the students more likely to answer the question correctly than a student not hearing the lectures.

Chapter 2: Using Surveys and Experimental Studies to Gather Data

2.9 **a.** Alumni (men only?) graduating from Yale in 1924.
b. No. Alumni whose addresses were on file 25 years later would not necessarily be representative of their class.
c. Alumni who *responded* to the mail survey would not necessarily be representative of those who were *sent* the questionnaires. Income figures may not be reported accurately (intentionally) or may be rounded off to the nearest $5,000, say, in a self-administered questionnaire.
d. Rounding income responses would make the figure $25,111 highly unlikely. The fact that higher-income respondents would be more likely to respond (bragging) and the fact that incomes are likely to be exaggerated would tend to make the estimate too high.

2.14 **a.** Heat treatment temperature and type of hardener
b. Heat treatment temperature: 175°F, 200°F, 225°F and 250°F
Type of hardener: H_1, H_2, H_3
c. Manufacuring plants
d. Plastic pipe
e. Locations on plastic pipe
f. 2 Pipes per treatment from each plant
g. None
h. 12 treatments

2.26 If phosphorus first: [P, N]

[10, 40], [10, 50], [10, 60], then [20, 60], [30, 60]
Or [20, 40], [20, 50], [20, 60], then [10, 60], [30, 60]
Or [30, 40], [30, 50], [30, 60], then [10, 60], [20, 60]

If nitrogen first: [N, P]

[40, 10], [40, 20], [40, 30], then [50, 30], [60, 30]
Or [50, 10], [50, 20], [50, 30], then [40, 30], [60, 30]
Or [60, 10], [60, 20], [60, 30], then [40, 30], [50, 30]

2.28 **a.** Group dogs by sex and age:

Group	Dog
Young female	2, 5, 13, 14
Young male	3, 5, 6, 16
Old female	5, 9, 10, 11
Old male	4, 8, 12, 15

b. Generate a random permutation of the numbers 1 to 16:
15 7 4 11 3 13 8 1 12 16 2 5 6 10 9 14
Go through the list and the first two numbers that appear in each of the four groups receive treatment L_1 and the other two receive treatment L_2.

Group	Treatment–Dog
Young female	$2-L_2, 7-L_1, 13-L_1, 14-L_2$
Young male	$3-L_1, 5-L_2, 6-L_2, 16-L_1$
Old female	$1-L_1, 9-L_2, 10-L_2, 11-L_1$
Old male	$4-L_1, 8-L_2, 12-L_2, 15-L_1$

***Expanded Answers to Selected Exercises are available at www.cengage.com/statistics/ott.**

Chapter 3: Data Description

3.4 a. Range = 1.05 − 0.72 = 0.33
b. Frequency histogram should be plotted with 7 classes ranging from 0.705 to 1.555.
The intervals have width .05.

3.7 a. Construct separate relative frequency histograms.
b. The histogram for the new therapy has one more class than the standard therapy. This would indicate that the new therapy generates a few more large values than the standard therapy. However, there is not convincing evidence that the new therapy generates a longer survival time.

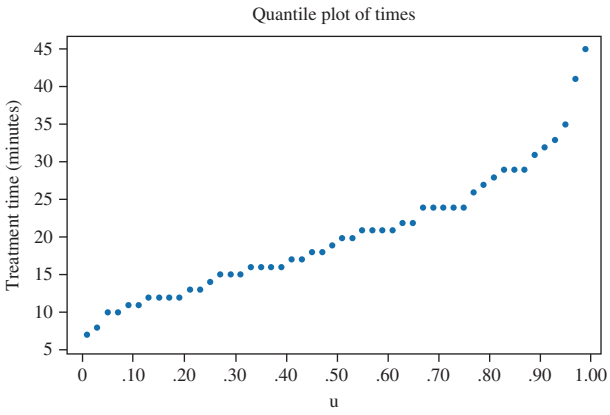
3.8 The plot has a bimodal shape. This would be an indication that there are two separate populations. However, the evidence is not very convincing because the individual plots were similar in shape with the exception that the New Therapy had a few times somewhat larger than the survival times obtained under the Standard Therapy.

3.12 The shapes of the 1985, 1996, and 2002 histograms and stem-and-leaf plots are asymmetric. The six plots are unimodal and left skewed.

3.15 Mean = 55.19, median = 58, two modes: 24, 58

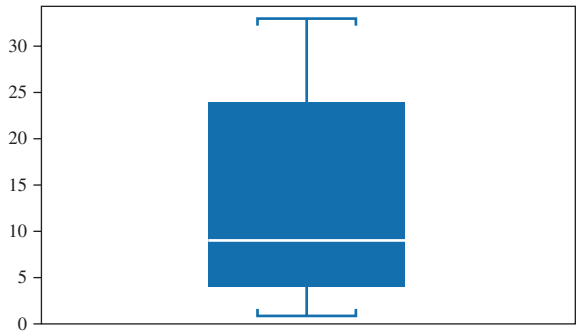
3.21 a. Mean = 8.04, median = 1.54
b. Terrestrial: mean = 15.01, median = 6.03
Aquatic: mean = .38, median = .375

3.29 The quantile plot is given at right.
a. The 25th percentile is the value associated with $u = .25$ on the graph, which is 14 minutes. Also, by definition, 14 minutes is the 25th percentile, since 25% of the times are less than or equal to 14 and 75% of the times are greater than or equal to 14 minutes.
b. Yes; the 90th percentile is 31.5 minutes. This means that 90% of the patients have a treatment time less than or equal to 31.5 minutes (which is less than 40 minutes).



ANSWER 3.29

3.33 The box plot is given at right.



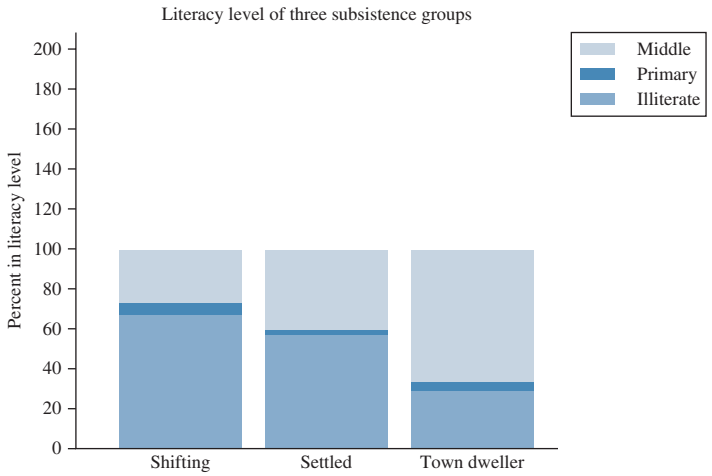
ANSWER 3.33

3.35 a. Can: $Q_1 \neq 1.45$, $Q_2 \neq 1.65$, $Q_3 \neq 2.4$
Dry: $Q_1 \neq .55$, $Q_2 \neq .60$, $Q_3 \neq .7$
b. Canned dog food is more expensive (median much greater than that for dry dog food), highly skewed to the right with a few large outliers. Dry dog food is slightly left skewed with a considerably smaller degree of variability than canned dog food.

3.39 a. The stacked bar graph is given at right.
b. Illiterate: 46%; primary schooling: 4%; at least middle school: 50%

Shifting cultivators: 28%; settled agriculturists: 21%; town dwellers: 51%

There is a marked difference in the distribution in the three literacy levels for the three subsistence groups. Town dwellers and shifting cultivators have the reverse trends in the three categories, whereas settled agriculturists fall into essentially two classes.

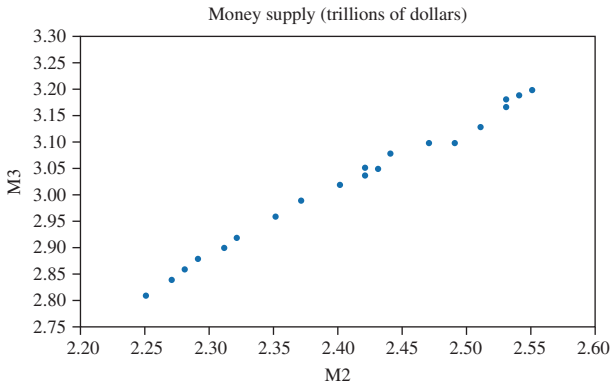


ANSWER 3.39

3.41 A scatterplot of M3 versus M2 is given at right.

a. Yes, it would because we want to determine the relative changes in the two over the 20 month period of time.

b. See scatterplot. The two measures follow an approximately increasing linear relationship.



ANSWER 3.41

Chapter 4: Probability and Probability Distributions

4.1 a. Subjective probability
b. Classical probability
c. Relative frequency

4.27 a. P(both customers pay in full) = (.70) (.70) = .49
b. P(at least one of two customers pays in full)
= 1 − P (neither customer pays in full)
= $1 - (1 - .70)(1 - .70) = 1 - (.30)^2 = .91$

4.29 Let D be the event loan is defaulted, R_1 applicant is poor risk, R_2 fair risk, and R_3 good risk.

$P(D) = .01$, $P(R_1|D) = .30$, $P(R_2|D) = .40$, $P(R_3|D) = .30$,
$P(\overline{D}) = .99$, $P(R_1|\overline{D}) = .10$, $P(R_2|\overline{D}) = .40$, $P(R_3|\overline{D}) = .50$

$$P(D|R_1) = \frac{P(R_1|D)P(D)}{P(R_1|D)P(D) + P(R_1|\overline{D})P(\overline{D})} = \frac{(.30)(.01)}{(.30)(.01) + (.10)(.99)} = .0294$$

4.31
$$P(D_1|A_1) = \frac{P(A_1|D_1)P(D_1)}{P(A_1|D_1)P(D_1) + P(A_1|D_2)P(D_2) + P(A_1|D_3)P(D_3) + P(A_1|D_4)P(D_4)}$$
$$= \frac{(.90)(.028)}{(.90)(.028) + (.06)(.012) + (.02)(.032) + (.02)(.928)} = .55851$$
$$P(D_2|A_2) = \frac{(.80)(.012)}{(.05)(.028) + (.80)(.012) + (.06)(.032) + (.01)(.928)} = .43243$$
$$P(D_3|A_3) = \frac{(.82)(.032)}{(.03)(.028) + (.05)(.012) + (.82)(.032) + (.02)(.928)} = .56747$$

4.33 Let F be the event fire occurs and T_i be the event a type i furnace is in the home for $i = 1, 2, 3, 4$, where T_4 represents other types.

$$P(T_1|F) = \frac{P(F|T_1)P(T_1)}{P(F|T_1)P(T_1) + P(F|T_1)P(T_1) + P(F|T_1)P(T_1) + P(F|T_1)P(T_1)}$$
$$= \frac{(.05)(.30)}{(.05)(.30) + (.03)(.25) + (.02)(.15) + (.04)(.30)} = .40$$

4.35
$$P(A_2|B_1) = \frac{(.17)(.15)}{(.08)(.25) + (.17)(.15) + (.10)(.12)} = .4435$$
$$P(A_2|B_2) = \frac{(.12)(.15)}{(.18)(.25) + (.12)(.15) + (.14)(.12)} = .2256$$
$$P(A_2|B_3) = \frac{(.07)(.15)}{(.06)(.25) + (.07)(.15) + (.08)(.12)} = .2991$$
$$P(A_2|B_4) = \frac{(.64)(.15)}{(.68)(.25) + (.64)(.15) + (.68)(.12)} = .2762$$

4.43 Yes, if the people not responding are ignored.

4.45 Binomial experiment with $n = 15$, $\pi = .2$, and y = number exceeding limit

a. $P(y = 15) \approx 0$
b. $P(y = 6) = .043$
c. $P(y \geq 6) = 1 - P(y < 6) = 1 - (P(0) + P(1) + P(2) + P(3) + P(4) + P(5)) = 1 - (.0389) = .0611$
d. $P(y = 0) = .0352$

4.73 No. The sample would be biased toward homes for which the homeowner is at home much of the time. For example, the sample would tend to include more people who work at home and retired persons.

4.75 Starting at column 2, line 1, we obtain 150, 465, 483, 930, 399, 069, 729, 919, 143, 368, 695, 409, 939, 611, 973, 127, 213, 540, 539, 976, 912, 584, 323, 270, 330. These would be the women selected for the study.

4.77 The sampling distribution would have a mean of 60 and a standard deviation of $\frac{5}{\sqrt{16}} = 1.25$. If the population distribution is somewhat mound shaped then the sampling distribution of $\bar{y}$ should be approximately mound shaped. In this situation, we would expect approximately 95% of the possible values of $\bar{y}$ to lie in $60 \pm (2)(1.25) = (57.5, 62.5)$.

4.83 $\mu = 2.1, \quad \sigma = .3$

a. $P(y > 2.7) = P\left(z > \dfrac{2.7 - 2.1}{0.3}\right) = P(z > 2) = .0228$

b. $P(z > .6745) = .25 \Rightarrow y_{.75} = 2.1 + (.6745)(.3) = 2.30$

c. Let μ_N be the new value of the mean. We need $P(y > 2.7) < .05$.
From Table 1 in the Appendix, $.05 = P(z > 1.645)$ and $.05 = P(y \le 2.7) = P\left(\dfrac{y - \mu_N}{.3} > \dfrac{2.7 - \mu_N}{.3}\right) \Rightarrow \dfrac{2.7 - \mu_N}{.3} = 1.645 \Rightarrow$
$\mu_N = 2.7 - (.3)(1.645) = 2.2065$.

4.85 Individual baggage weight has $\mu = 95; \sigma = 35$; total weight has mean $n\mu = (200)(95) = 19{,}000$;
and standard deviation $\sqrt{n}\sigma = \sqrt{200}(35) = 494.97$. Therefore, $P(y > 20{,}000) = P\left(z > \dfrac{20{,}000 - 19{,}000}{494.97}\right) = P(z > 2.02) = .0217$.

4.89 $n = 10, \pi = .5$

a. $P(4 \le y \le 6) = P(y = 4) + P(y = 5) + P(y = 6) = \binom{10}{4}(.5)^4(.5)^6 + \binom{10}{5}(.5)^5(.5)^5 + \binom{10}{6}(.5)^6(.5)^4 = .65625$

b. $\mu = (10)(0.5) = 5; \sigma = \sqrt{(10)(0.5)(0.5)} = 1.58$;

$P(4 \le y \le 6) = P\left(z < \dfrac{6 - 5}{1.58}\right) - P\left(z < \dfrac{4 - 5}{1.58}\right) = P(z < .63) - P(z < -.63) = .4714$. It did not work well.

4.99 No, there is strong evidence that the new fabric has a greater mean breaking strength.

4.101 $\mu = 5.35, \quad \sigma = .12$

a. $P(y > \log(250)) = P(y > 5.52) = P\left(z > \dfrac{5.52 - 5.35}{.12}\right) = .0078 = .78\%$

b. $P(\log(150) < y < \log(250)) = P(5.01 < y < 5.52) = P\left(\dfrac{5.01 - 5.35}{.12} < z < \dfrac{5.52 - 5.35}{.12}\right) = .9194 = 91.94\%$

c. $P(y > \log(300)) = P(y > 5.7) = P\left(z > \dfrac{5.7 - 5.35}{.12}\right) = .0018 = .18\%$

4.103 $n = 20{,}000, \pi = .0001$. There are two possible outcomes, and each birth is an independent event. We cannot use the normal approximation because $n\pi = (20{,}000)(.0001) = 2 < 5$. We can use the binomial formula:
$P(y \ge 1) = 1 - P(y = 0) = 1 - \binom{20{,}000}{0}(.0001)^0(.9999)^{20{,}000} = .8647$

Chapter 5: Inferences About Population Central Values

5.13 $\hat{\sigma} = 13, E = 3, \alpha = .01 \Rightarrow n = \dfrac{(2.58)^2(13)^2}{(3)^2} = 125$

5.21 $H_0: \mu \le 2$ versus $H_a: \mu > 2, \bar{y} = 2.17, s = 1.05, n = 90$

a. $z = \dfrac{2.17 - 2}{1.05/\sqrt{90}} = 1.54 < 1.645 = z_{0.05} \Rightarrow$

Fail to reject H_0. The data do not support the hypothesis that the mean has been increased from 2.

b. $\beta(2.1) = P(z \le 1.645 - \dfrac{|2 - 2.1|}{1.05/\sqrt{90}} = P(z \le .74) = .7704$

5.24 $n = \dfrac{(80)^2(1.645 + 1.96)^2}{(525 - 550)^2} = 133.1 \Rightarrow n = 134$

5.27 $H_0: \mu \le 30$ versus $H_a: \mu > 30$,
$\alpha = .05, n = 37, \bar{y} = 37.24, s = 37.12$

a. $z = \dfrac{37.24 - 30}{37.12/\sqrt{37}} = 1.19 < 1.645 = z_{0.05} \Rightarrow$ Fail to reject H_0.

There is not sufficient evidence to conclude that the mean lead concentration exceeds 30 mg kg^{-1} dry weight.

b. $\beta(50) = P\left(z \le 1.645 - \dfrac{|30 - 50|}{37.12/\sqrt{37}}\right) = P(z \le -1.63) = .0513$

c. No, the data values are not very close to the straight line in the normal probability plot.

d. No; since there is a substantial deviation from a normal distribution, the sample size should be somewhat larger to use the z test. Section 5.8 provides an alternative test statistic for handling this situation.

5.33 $H_0: \mu = 1.6$ versus $H_a: \mu \ne 1.6$,
$n = 36, \bar{y} = 2.2, s = .57, \alpha = .05$

$p\text{-value} = 2P\left(z \ge \dfrac{|2.2 - 1.6|}{.57/\sqrt{36}}\right) = 2P(z \ge 6.32) < .0001 < .05 = \alpha \Rightarrow$

Yes, there is significant evidence that the mean time delay differs from 1.6 seconds.

5.39 $n = 15, \bar{y} = 31.47, s = 5.04$

a. $31.47 \pm (2.977)(5.04)/\sqrt{15} \Rightarrow 31.47 \pm 3.87 \Rightarrow (27{,}600, 35{,}340)$ is a 99% C.I. on the mean miles driven.

b. H_0: $\mu \geq 35$ versus H_a: $\mu < 35$;

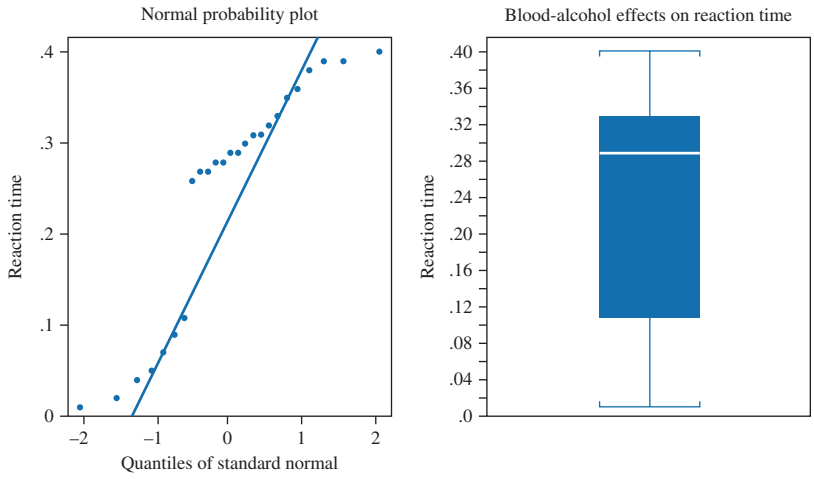
$$t = \frac{31.47 - 35}{5.04/\sqrt{15}} = -2.71 \Rightarrow \text{Reject } H_0 \text{ if } t \leq -2.624.$$

Reject H_0 and conclude the data support the hypothesis that the mean miles driven is less than 35,000 miles. Level of significance is given by p-value $= P(t \leq -2.71) \Rightarrow .005 < p\text{-value} < .01$.

5.41 **a.** $4.95 \pm (2.365)(0.45)/\sqrt{8} \Rightarrow 4.95 \pm .38 \Rightarrow (4.57, 5.33)$ is a 95% C.I. on the mean dissolved oxygen level.

b. There is inconclusive evidence that the mean is less than 5 ppm since the C.I. contains values both less and greater than 5 ppm.

c. H_0: $\mu \geq 5$ versus H_a: $\mu < 5$, p-value $= P(t \leq -.31) \Rightarrow .25 < p\text{-value} < .40$ (using a computer program, p-value $= .3828$). Fail to reject H_0 and conclude the data do not support that the mean is less than 5 ppm.



ANSWER 5.52

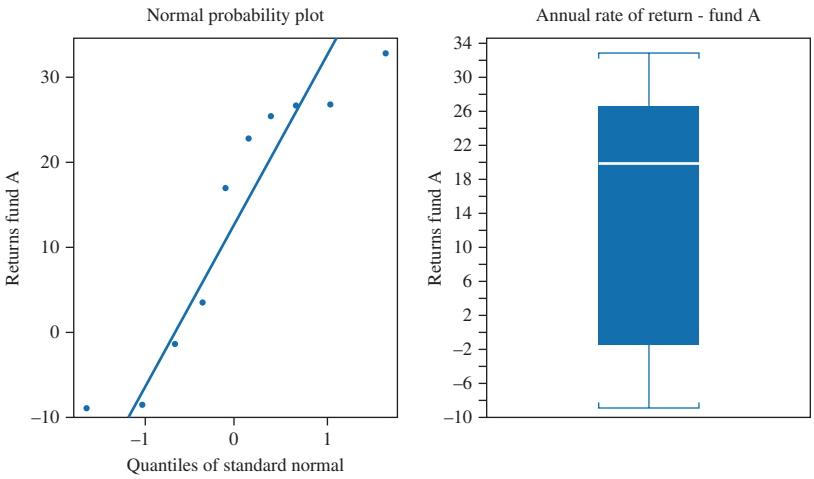
5.52 **a.** The graphs are given at right.

b. 99% C.I. on mean: $.247 \pm (2.979)(.129)/\sqrt{25} \Rightarrow (.175, .319)$; 99% C.I. on median: $(y_{(5)}, y_{(21)}) \Rightarrow (.07, .36)$

c. Yes, $t = \dfrac{.247 - 0}{.129/\sqrt{25}} = 9.57 \Rightarrow p\text{-value} = P(t \geq 9.57) < .0001$. Thus, there is significant evidence of an increase in mean reaction time.

d. Yes, $B = 25 > 21 \Rightarrow$ Reject H_0 at the $\alpha = .001$ level. Thus, there is significant evidence of an increase in median reaction time.

e. Using the normal probability plot and boxplot, it is observed that the data appear to be from a distribution that is bimodal, skewed to the left. Thus, the median is a more appropriate representative of reaction time differences.

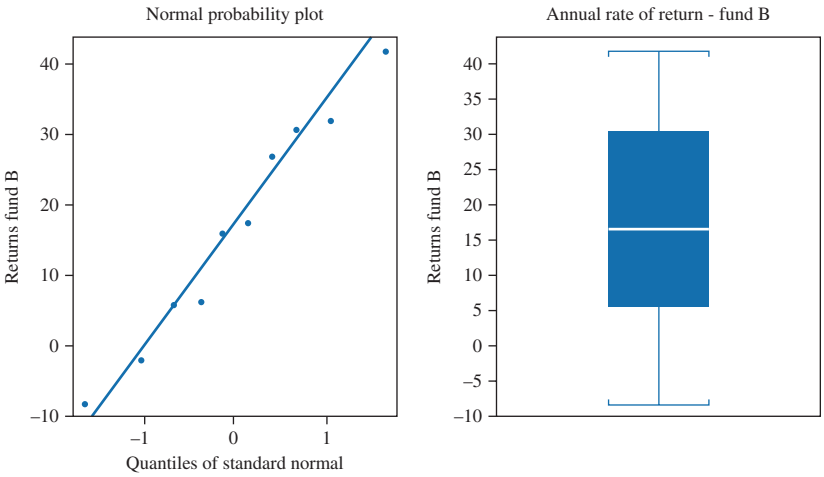


5.54 **a.** Fund A: 95% C.I. on the mean: $13.65 \pm (2.262)(15.87)/\sqrt{10} \Rightarrow (2.30, 25.00)$, median $= 20$; 95% C.I. on the median: $(y_{(2)}, y_{(9)}) \Rightarrow (-8.5, 26.7)$

Fund B: 95% C.I. on the mean: $16.56 \pm (2.262)(16.23)/\sqrt{10} \Rightarrow (4.95, 28.17)$, median $= 16.6$; 95% C.I. on the median: $(y_{(2)}, y_{(9)}) \Rightarrow (-2.1, 31.9)$

b. The normal probability and box plots are given at right.

Based on the boxplots and normal probability plots, the median is the more appropriate measure for fund A, and the mean is more appropriate for fund B.



ANSWER 5.54

5.62 **a.** $\bar{y} = 74.2$; 95% C.I.: $74.2 \pm (2.145)(44.2)/\sqrt{15} \Rightarrow (49.72, 98.68)$

b. H_0: $\mu \leq 50$ versus H_a: $\mu > 50$,
$n = 15, \alpha = .05$

$$p\text{-value} = P\left(t \geq \frac{74.2 - 50}{44.2/\sqrt{15}}\right) = P(t \geq 2.12)$$
$$= .0262 < .05 = \alpha$$

Yes, there is sufficient evidence to conclude that the average daily output is greater than 50 tons of ore.

5.64 **a.** The summary statistics are given here:

Time	Mean	Std Dev	*n*	95% C.I.
6 A.M.	.128	.0355	15	(.108, .148)
2 P.M.	.116	.0406	15	(.094, .138)
10 P.M.	.142	.0428	15	(.118, .166)
All day	.129	.0403	45	(.117, .141)

b. No, the three C.I.s have a considerable overlap.

c. $H_0: \mu \geq .145$ versus $H_a: \mu < .145$

$$p\text{-value} = P\left(t \leq \frac{.129 - .145}{.0403/\sqrt{45}}\right) = P(t \leq -2.66) = .0054$$

There is significant evidence (very small p-value) that the average SO_2 level using the new scrubber is less than .145.

5.66 $n = \dfrac{(12.36)^2(1.96)^2}{(1)^2} = 586.9 \Rightarrow n = 587$

5.68 $n = 40, \bar{y} = 58, s = 10$

99% C.I. on μ: $58 \pm (2.708)(10)/\sqrt{40} \Rightarrow (53.7, 62.3)$

5.76 **a.** Let $\mu_C = \mu_{Before} - \mu_{After}$. The probabilities of Type II error are computed using Table 3 in the Appendix with $d = \dfrac{|\mu_C - 0|}{7.54}$ and are given here:

μ_C	−1.0	−2.0	−3.0	−4.0	−5.0	−6.0	−7.0	−8.0	−9.0
d	0.13	.27	.40	0.53	0.66	.80	.93	1.06	1.19
$\beta(\mu_C)$	0.89	0.81	0.68	0.54	0.39	0.25	0.14	0.07	0.03

The probabilities of Type II error are large for values of μ_C, which are of practical importance.

b. Since the probabilities of Type II errors are large, the sample size should be increased. The models, ages, and conditions of the cars used in the study should be considered. The type of driving conditions and experience of drivers are also important factors to be considered in order for the results to be generalizable to a broad population of potential users of the device.

Chapter 6: Inferences Comparing Two Population Central Values

6.5 **a.** H_0: $\mu_{26} - \mu_5 \geq 0$ versus H_a: $\mu_{26} - \mu_5 < 0$; reject H_0 if $t \leq -1.812$.

$$t = \frac{165.8 - 378.5}{19.9\sqrt{\frac{1}{6} + \frac{1}{6}}} = -18.51 < -1.812 \Rightarrow$$ Reject H_0 and conclude there is significant evidence that μ_{26} is less than μ_5, with p-value $< .0005$.

b. The sample sizes are too small to evaluate the normality condition, but the sample variances are fairly close, considering the sample sizes. We would need to check with the experimenter to determine if the two random samples were independent.

c. A 95% C.I. on the mean difference is (−238.3, −187.1), which indicates that the average warm temperature rat blood pressure is between 187 and 239 units lower than the average 5°C rat blood pressure.

6.7 **a.** H_0: $\mu_U \leq \mu_S$ versus H_a: $\mu_U > \mu_S$; p-value $< .0005 \Rightarrow$ The data provide sufficient evidence to conclude that successful companies have a lower percentage of returns than unsuccessful companies.

b. $n_1 + n_2 - 2 = 98 =$ df for pooled t test. The printout shows df = 86, which is the df for the separate variance test.

c. The boxplots indicate that both data sets appear to be from normally distributed populations; however, the successful data sets indicate a higher variability than the unsuccessful.

d. A 95% C.I. on the difference in the mean percentages is (2.15%, 4.35%) $\Rightarrow$ We are 95% confident that successful businesses have roughly 2% to 4.5% fewer returns.

6.11 **a.** H_0: $\mu_{96} \geq \mu_{82}$ versus H_a: $\mu_{96} < \mu_{82}$;

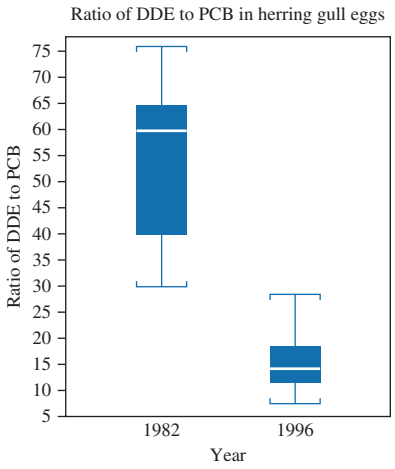
$$t = \frac{15.52 - 54.30}{\sqrt{\frac{(5.96)^2}{13} + \frac{(15.7)^2}{13}}} = -8.35 \Rightarrow \text{with df} = 13, p\text{-value} < .0005 \Rightarrow$$

Reject H_0 and conclude the data provide sufficient evidence that there has been a significant decrease in mean PCB content.

b. A 95% C.I. on the difference in the mean PCB contents of herring gull eggs is $(-48.7, -28.9)$, which would indicate that the decrease in mean PCB content from 1982 to 1996 is between 28.9 and 48.7.

c. The boxplots are given at right.

The boxplots of the PCB data from the two years both appear to support random samples from normal distributions, although the 1982 data are somewhat skewed to the left. The variances for the two years are substantially different; hence, the separate variance t test was applied in part (a).



ANSWER 6.11c

d. Since the data for 1982 and 1996 were collected at the same sites, there may be correlation between the two years. There may also be spatial correlation depending on the distance between sites.

6.27 **a.** To conduct the study using independent samples, the 30 participants should be very similar relative to age, body fat percentage, diet, and general health prior to the beginning of the study. The 30 participants would then be randomly assigned to the two treatments.

b. The participants should be matched to the greatest extent possible based on age, body fat, diet, and general health before the treatment is applied. Once the 15 pairs are configured, the two treatments are randomly assigned within each pair of participants.

c. If there is a large difference in the participants with respect to age, body fat, diet, and general health and if the pairing results in a strong positive correlation in the responses from paired participants, then the paired procedure would be more effective. If the participants are quite similar in the desired characteristics prior to the beginning of the study, then the independent samples procedure would yield a test statistic having twice as many df as the paired procedure and hence would be more powerful.

6.35 **a.** The boxplot and normal probability plots both indicate that the distribution of the data is somewhat skewed to the left. Hence, the Wilcoxon would be more appropriate, although the paired t test would not be inappropriate, since the differences are nearly normal in distribution.

b. H_0: The distribution of differences (female minus male) is symmetric about 0 versus H_a: The differences (female minus male) tend to be larger than 0.
With $n = 20, \alpha = .05, T = T_-$, reject H_0 if $T_- \leq 60$.
From the data, we obtain $T_- = 18 < 60$; thus, reject H_0 and conclude that repair costs are generally higher for female customers.

6.43 **a.** H_0: $\mu_{Narrow} = \mu_{Wide}$ versus H_a: $\mu_{Narrow} \neq \mu_{Wide}$;

$$t = \frac{118.37 - 110.20}{\sqrt{\frac{(7.87)^2}{12} + \frac{(4.71)^2}{15}}} = 3.17 \Rightarrow \text{with df} \approx 17, .002 < p\text{-value} < .010 \Rightarrow$$

Reject H_0 and conclude there is sufficient evidence in the data that the two types of jets have different average noise levels.

b. A 95% C.I. on $\mu_{Wide} - \mu_{Narrow}$ is $(2.73, 13.60)$.

c. Because maintenance could affect noise levels, jets of both types from several different airlines and manufacturers should be selected. They should be of approximately the same age. This study could possibly be improved by pairing narrow and wide body airplanes based on factors that may affect noise level.

6.47 **a.** H_0: $\mu_{Within} = \mu_{Out}$ versus H_a: $\mu_{Within} \neq \mu_{Out}$

Since both n_1 and n_2 are greater than 10, the normal approximation can be used.
$T = 122, \mu_T = (12)(12 + 14 + 1)/2 = 162, \sigma = \sqrt{(12)(14)(12 + 14 + 1)/12} = 19.44$

$$z = \frac{122 - 162}{19.44} = 2.06 \Rightarrow p\text{-value} = .0394 \Rightarrow$$

Reject H_0 and conclude the data provide sufficient evidence that there is a difference in average population abundance.

b. The Wilcoxon rank sum test requires independently selected random samples from two populations that have the same shape but may be shifted from one another.

c. The two population distributions may have different variances but the Wilcoxon rank sum test is very robust to departures from the required conditions.

d. The separate variance test failed to reject H_0 with a p-value of .384. The Wilcoxon test rejected H_0 with a p-value of .0394. The difference in the two procedures is probably due to the skewness observed in the outside data set. This can result in inflated p-values for the t test, which relies on a normal distribution when the sample sizes are small.

6.51 **a.** H_0: $\mu_{Low} = \mu_{Con}$ versus H_a: $\mu_{Low}, \neq \mu_{Con}$
Separate variance t test: $t = -2.09$ with df ≈ 35, p-value $= .044$. $\Rightarrow$
Reject H_0 and conclude there is significant evidence of a difference in the mean drop in blood pressure between the low-dose and control groups.

b. 95% C.I. on $\mu_{Low} - \mu_{Con}$: $(-51.3, -0.8)$; that is, the low-dose group's mean drop in blood pressure was, with 95% confidence, 51.3 to .8 points less than the mean drop observed in the control group.

c. Provided the researcher independently selected the two random samples of participants, the conditions for using a pooled t test were satisfied, since the plots do not detect a departure from a normal distribution and the sample variances are similar in size.

6.57 Let d = before − after

a. H_0: $\mu_{Before} = \mu_{After}$ versus H_a: $\mu_{Before} \neq \mu_{After}$;

$$t = \frac{-0.122}{0.106/\sqrt{15}} = -4.45 \text{ with df} = 14, p\text{-value} < .0005. \Rightarrow$$

Reject H_0 and conclude the data provide sufficient evidence that the mean soil pH has changed after mining on the land.

b. H_a: $\mu_{Before} \neq \mu_{After}$

c. 99% C.I. on $\mu_{Before} - \mu_{After}$: (.04, .20)

d. The findings are highly significant (p-value < .0005), statistically. The question is, How significant are the results in a practical sense? Unless a change in pH of between .04 and .20 has an impact on the soil with respect to common usages of the soil, the mining company should not be cited.

6.59 **a.** The average potency after 1 year is different than the average potency right after production.

b. The two test statistics are equal, since the sample sizes are equal: $t = t' = 4.2368$.

c. The p-values are different, since the test statistics have different degrees of freedom (df): for t, p-value = .0006, and for t', p-value = .0005.

d. In this particular experiment, the test statistics reach the same conclusion, reject H_0.

e. Because $s_1 \approx s_2$ and a test of equal variances has p-value equal to .3917, the pooled t test (t) would be the more appropriate test statistic.

Chapter 7: Inferences About Population Variances

7.7 **a.** The middle 50% of the data are symmetric, but there are four outliers. Since the sample size is 150, a few outliers would be expected. However, 4 out of 150 may indicate the population distribution may have heavier tails than a normal distribution. This may cause the values of the sample standard deviation to be inflated.

b. 99% C.I. on σ: $\left(\sqrt{\frac{(150-1)(9.537)^2}{197.21}}, \sqrt{\frac{(150-1)(9.537)^2}{108.29}}\right) \Rightarrow (8.290, 11.187)$

c. H_0: $\sigma^2 \leq 90$ versus H_a: $\sigma^2 > 90$

With $\alpha = .05$, reject H_0 if $\frac{(n-1)(s)^2}{90} \geq 178.49$.

$\frac{(150-1)(9.537)^2}{90} = 150.58 < 178.49 \Rightarrow$

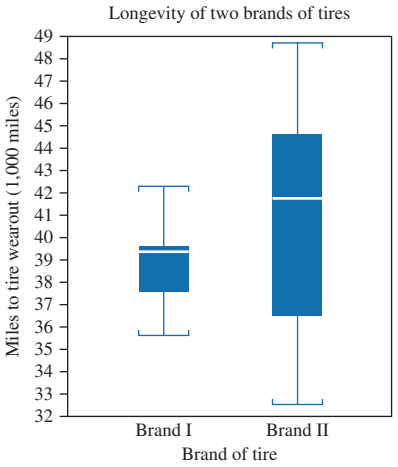
Fail to reject H_0 and conclude the data fail to support the statement that σ^2 is greater than 90.

7.19 The skewness in the data produces outliers, which may greatly distort both the mean and the standard deviation. Thus, BFL's test statistic minimizes both of these effects by replacing the mean with the median and using the absolute deviations about the median in place of the squared deviations about the mean.

7.20 **a.** The boxplots are at right.
The boxplots and normal probability plots indicate that both samples are from normally distributed populations.

b. The C.I.s are given here:

Method	n	Mean	95% C.I. on μ	Std. Dev.	95% C.I. on σ
I	10	38.79	(37.39, 40.19)	1.9542	(1.34, 3.57)
II	10	40.67	(36.68, 44.66)	5.5791	(3.84, 10.19)



ANSWER 7.20

c. A comparison of the population variances yields:

H_0: $\sigma_I^2 = \sigma_{II}^2$ versus H_a: $\sigma_I^2 \neq \sigma_{II}^2$

With $\alpha = .01$, reject H_0 if $\frac{s_I^2}{s_{II}^2} \leq \frac{1}{6.54} = .15$ or $\frac{s_1^2}{s_2^2} \geq 6.54$.

$s_1^2/s_2^2 = (5.5791)^2/(1.9542)^2 = 8.15 > 6.54 \Rightarrow$

Reject H_0 and conclude there is significant evidence that the population variances are different.

A comparison of the population means using the separate variance t test yields:

H_0: $\mu_I = \mu_{II}$ versus H_a: $\mu_I \neq \mu_{II}$;

$$t = \frac{38.79 - 40.67}{\sqrt{\frac{(1.9542)^2}{10} + \frac{(5.5791)^2}{10}}} = -1.01 \text{ with df} = 11 \Rightarrow p\text{-value} = .336 \Rightarrow$$

Fail to reject H_0 and conclude that the data do not support a difference in the tread wear means for the two brands of tires. However, Brand I has a more uniform tread wear, as reflected by its significantly lower standard deviation.

7.22 **a.** H_0: $\sigma_1^2 \geq \sigma_2^2$ versus H_a: $\sigma_1^2 < \sigma_2^2$

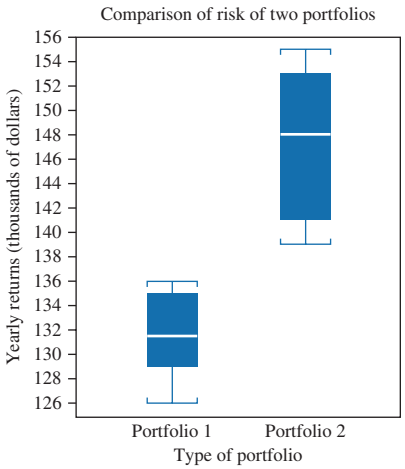
With $\alpha = .05$, reject H_0 if $\frac{s_2^2}{s_1^2} \geq 3.18$.

$s_2^2/s_1^2 = (5.9591)^2/(3.5963)^2 = 2.75 < 3.18 \Rightarrow$
Fail to reject H_0 and conclude there is not significant evidence that portfolio 2 has a larger variance than portfolio 1.

$$95\%\text{ C.I. on } \frac{\sigma_2^2}{\sigma_1^2}\text{: } \left(\frac{(5.9591)^2}{(3.5963)^2}(.248), \frac{(5.9591)^2}{(3.5963)^2}(4.03)\right) \Rightarrow (.68, 11.07)$$

b. p-value $= P(F_{(9,9)} \geq 2.75) \Rightarrow .05 < p\text{-value} < .10$

c. The boxplots are given at right.
From the boxplots, the condition of normality appears to be satisfied for both portfolios.



ANSWER 7.22c

7.24 The boxplots are given at right.
The boxplots indicate that both samples are from populations that are normally distributed but that have different levels of variability.
The C.I.s are given here:

Preparation	n	Mean	95% C.I. on μ	St. Dev.
A	13	27.62	(21.68, 33.55)	9.83
B	13	34.69	(32.26, 37.13)	4.03

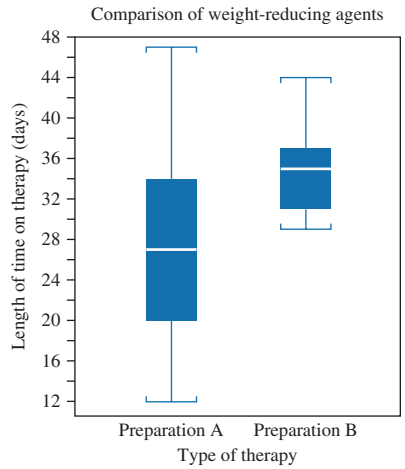
A comparison of the population variances yields:
H_0: $\sigma_A^2 = \sigma_B^2$ versus H_a: $\sigma_A^2 \neq \sigma_B^2$;
$s_A^2/s_B^2 = (9.83)^2/(4.03)^2 = 5.955 \Rightarrow .001 < p\text{-value} < .005 \Rightarrow$
Reject H_0 and conclude there is significant evidence that the population variances are different.

A comparison of the population means using the separate variance t test yields:

$$H_0\text{: } \mu_A = \mu_B \text{ versus } H_a\text{: } \mu_A \neq \mu_B;$$

$$t = \frac{27.62 - 34.69}{\sqrt{\frac{(9.83)^2}{13} + \frac{(4.03)^2}{13}}} = -2.40 \text{ with df} = 15 \Rightarrow p\text{-value} = .030 \Rightarrow$$

Reject H_0 and conclude that the data indicate a difference in the mean length of time people remain on the two therapies.



ANSWER 7.24

Chapter 8: Inferences About More Than Two Population Central Values

8.7 **a.** The AOV F test yields $F = \dfrac{2.292/2}{11.738/18} = 2.06 < 3.55 = F_{.05,\,2,\,18}$.
Thus, there is not significant evidence of a difference in the mean soil densities for the three grazing regimens.

b. The associated p-value is p-value $= 1 - pf(2.06,\ 2,18) = .156 > .05$, thus confirming our conclusion in part (a).

c. Based on a plot of the residuals versus the fitted values, the condition of constant variance does not appear to be violated. This is confirmed by the BFL test, which has a p-value equal to 0.366.
The normal quantile plot of the residuals indicates somewhat of a deviation from normality with the test for normality having p-value $= 0.049$. Based on the robustness of the F text with modest deviations from normality, the p-value from the F test will be considered valid.

8.27 **a.** The BFL test yields $L = 2.74$ with $p = .042$. Thus, there is significant evidence that the equal variance condition is also violated.

b. The AOV table is given here:

Source	df	SS	MS	F	p-value
Supplier	4	28,024	7,006.09	265.94	$< .0001$
Error	40	1,054	26.34		
Total	44	29,087			

With p-value $< .0001$, reject H_0 and conclude there is a significant evidence of a difference in the mean deviations of the five suppliers.

c. The Kruskal–Wallis test yields $H = 41.59$ with df $= 2 \Rightarrow p$-value $< .0001$. Thus, reject H_0 and conclude there is a significant difference in the distributions of deviations for the five suppliers.

8.29 a. Based on the boxplots and the normal probability plot, the condition of normality of the population distributions appears to be satisfied.
The BFL test yields $L = .17$ with p-value $= .913 \Rightarrow$ There is not significant evidence of a difference in the four population variances.

b. From the AOV table, we have p-value $< .001$. Thus, there is significant evidence that the mean ratings differ for the four groups.

c. 95% C.I. on μ_I : $8.3125 \pm 2.048\dfrac{\sqrt{.9763}}{\sqrt{8}} = (7.6, 9.0)$

95% C.I. on μ_{II} : $6.4375 \pm 2.048\dfrac{\sqrt{.9763}}{\sqrt{8}} = (5.7, 7.1)$

95% C.I. on μ_{III} : $4.0000 \pm 2.048\dfrac{\sqrt{.9763}}{\sqrt{8}} = (3.3, 4.7)$

95% C.I. on μ_{IV} : $2.5000 \pm 2.048\dfrac{\sqrt{.9763}}{\sqrt{8}} = (1.8, 3.2)$

8.31 a. The model for this experiment is given by

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}; \quad i = 1, 2, 3 \quad \text{and} \quad j = 1, \cdots, n_i$$

where $n_1 = 12, n_2 = 14, n_3 = 11$; μ = overall mean; τ_i = effect of ith division; ε_{ij} = random error associated with the jth response from the ith division.

$F = \dfrac{286.3/2}{611.1/34} = 7.97$ with p-value $= 1 - pf(7.97, 2, 34) = .0015 < .01 \Rightarrow$

There is significant evidence of a difference in the mean responses for the three divisions.

8.33 The Kruskal–Wallis test yields $H' = 16.56$ with df $= 3 \Rightarrow p$-value $< .001$.
There is significant evidence of a difference in the distribution of the yields for the four varieties.
The two procedures yield similar conclusions.

8.35 a. $F = \dfrac{4{,}020.0/3}{881.9/36} = 54.70$ with df $= 3, 36 \Rightarrow p$-value $< .001 < .05 \Rightarrow$

There is significant evidence of a difference in the average leaf sizes under the four growing conditions.

b. 95% C.I. on μ_A : $23.37 \pm 2.028\dfrac{\sqrt{881.9/36}}{\sqrt{10}} = (20.20, 26.54)$

95% C.I. on μ_B : $8.58 \pm 2.028\dfrac{\sqrt{881.9/36}}{\sqrt{10}} = (5.41, 11.75)$

95% C.I. on μ_C : $14.93 \pm 2.028\dfrac{\sqrt{881.9/36}}{\sqrt{10}} = (11.76, 18.10)$

95% C.I. on μ_D : $35.35 \pm 2.028\dfrac{\sqrt{881.9/36}}{\sqrt{10}} = (32.18, 38.52)$

The C.I. for the mean leaf size for condition D implies that the mean is much larger for condition D than for the other three conditions.

c. $F = \dfrac{18.08/3}{103.17/36} = 2.10$ with df $= 3, 36 \Rightarrow .05 < .10 < p$-value $< .25 \Rightarrow$

There is not significant evidence of a difference in the average nicotine contents under the four growing conditions.

d. From the given data, it is not possible to conclude that the four growing conditions produce different average nicotine contents.

e. No. If the testimony was supported by this experiment, then the test conducted in part (c) would have had the opposite conclusion.

8.37 a. Generate a plot for each diet.

b. The summary statistics are given here:

Diet	n	Mean	Variance
Control	6	3.783	0.278
Control + level 1 of A	6	5.500	0.752
Control + level 2 of A	6	6.983	0.334
Control + level 1 of B	6	7.000	0.128
Control + level 2 of B	6	9.383	0.086

c. The BFL test yields $L = 2.23$ with p-value $= .095 \Rightarrow$ There is not significant evidence of a difference in the five variances. The boxplots do not reveal any deviations from the normality condition.

d. $F = \dfrac{103.04/4}{7.885/25} = 81.67$ with df $= 4, 25 \Rightarrow p$-value $< .001 < .05 \Rightarrow$

There is significant evidence of a difference in the average weight gains under the five diets.

8.39 $F = \dfrac{1{,}146.33/2}{219.67/15} = 39.14$ with df $= 2, 15 \Rightarrow p\text{-value} < .001 < .05 \Rightarrow$

There is significant evidence of a difference in the average seedling heights for the three groups.

8.41 The value of the Kruskal–Wallis statistic is identical to the value calculated prior to replacing 9.8 with 15.8. This will not happen in general, but 9.8 was the largest value in the original data, and, hence, its rank would not be altered by increasing its size. If there is an extreme value in the data set, it may greatly alter the conclusion reached by the AOV F test. The Kruskal–Wallis test is not sensitive to extreme values, since it just replaces these extremes with their corresponding ranks.

8.43 The Kruskal–Wallis test yields identical results for the transformed and original data because the transformation was strictly increasing, which maintains the order of the data after the transformation has been performed.

$H = 9.89$ with df $= 2 \Rightarrow .005 < p\text{-value} < .01 < .05$ using the chi-square table

Thus, our conclusion is the same as was reached using the transformed data.

Chapter 9: Multiple Comparisons

9.5 **a.** $l_1 = 4\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5$

b. $l_2 = 3\mu_2 - \mu_3 - \mu_4 - \mu_5$

c. $l_3 = \mu_3 - 2\mu_4 + \mu_5$

d. $l_4 = \mu_3 - \mu_5$

9.13 The boxplot indicates the distribution of the residuals is slightly right skewed. This is confirmed with an examination of the normal probability plot. The BFL test yields $L = .24$ with p-value $= .917$. Thus, the conditions needed to run the AOV F test appear to be satisfied. From the output, $F = 15.68$ with p-value $< .0001 < .05$. Thus, we reject H_0 and conclude there is significant evidence of a difference in the average weight losses obtained using the five different agents.

9.17 **a.** $l_1 = \mu_{A_1} + \mu_{A_2} + \mu_{A_3} + \mu_{A_4} - 4\mu_S$

b. $l_2 = \mu_{A_1} - \mu_{A_2} + \mu_{A_3} - \mu_{A_4}$

c. $l_3 = \mu_{A_1} + \mu_{A_2} - \mu_{A_3} - \mu_{A_4}$

d. $l_4 = \mu_{A_1} + \mu_{A_3} - 2\mu_S$

9.20 **a.** Using Dunnett's procedure: $D = (1.94)\sqrt{2(52.62)/30} = 3.63$.

$\bar{y}_1 - \bar{y}_C = 5.2 > 3.63, \quad \bar{y}_2 - \bar{y}_C = 4.3 > 3.63 \Rightarrow$

There is significant evidence that both μ_1 and μ_2 are larger than μ_C.

b. Because the goal of the study was to determine if the use of herbicides increased the mean yield, the appropriate procedure would be one-sided.

c. There is significant evidence that both herbicides have larger mean yields than the control.

Chapter 10: Categorical Data

10.1 **b.** $n = 35, \hat{\pi} = .80 \Rightarrow y = 28 \Rightarrow \tilde{y} = 28 + .5(2.576)^2 = 31.318, \tilde{n} = 35 + (2.576)^2 = 41.636, \tilde{\pi} = 31.318/41.636 = .7522 \Rightarrow$ 99% C.I. is $.7522 \pm 2.576\sqrt{.7522(1 - .7522)/41.636} = (.580, .925)$. Without correction C.I. is $.8 \pm 2.576)\sqrt{(.8)(1 - .8)/35} = (.626, .974)$.

10.10 **a.** By grouping the classes into similar types, it might be possible to summarize the data more concisely. Percentages are helpful but would not add to 100% because one adult might use more than one of the remedies. The numerator of the percentage would refer to users of an OTC remedy and the denominator to the number of patients.

b. A 95% C.I. using the normal approximation requires that both $n\hat{\pi}$ and $n(1 - \hat{\pi})$ exceed 5. This condition would hold in every OTC category except room vaporizers and nasal sprays.

10.35 H_0: $\pi_1 = .0625, \pi_2 = .25, \pi_3 = .375, \pi_4 = .25, \pi_5 = .0625$

H_a: At least one of the π_is differs from its hypothesized value.

$E_i = n\pi_{io} \Rightarrow E_1 = 125(.0625) = 7.8125, E_2 = 125(.25) = 31.25,$

$E_3 = 125(.375) = 46.875, E_4 = 125(.25) = 31.25, E_5 = 125(.0625) = 7.8125$

$$\chi^2 = \sum_{i=1}^{5} \frac{(n_i - E_i)^2}{E_i} = 7.608 \text{ with df} = 5 - 1 = 4 \Rightarrow p\text{-value} > .107 \Rightarrow$$

Fail to reject H_0. The data appear to fit the hypothesized theory that the securities analysts perform no better than chance, however, we have no indication of the probability of a Type II error.

10.39 **a.** From the data, $\bar{y} = \dfrac{1}{100}\sum_i (n_i)(y_i) = 5.57$.

$$s^2 = \frac{1}{99}\sum_i n_i(y_i - 5.57)^2 = 1{,}056.5/99 = 10.67$$

b. Using $\mu = 5.5$, the Poisson table yields the following probabilities after combining the first two categories and combining the last four categories, so that $E_i > 1$ and only one E_i is less than 5:

k	≤ 1	2	3	4	5	6	7	8	≥ 9
$\pi_i = P(y = k)$	.0266	.0618	.1133	.1558	.1714	.1571	.1234	.0849	.1057
$E_i = 100\pi_i$	2.66	6.18	11.33	15.58	17.14	15.71	12.34	8.49	10.57

$\chi^2 = \sum_{i=1}^{9} \frac{(n_i - E_i)^2}{E_i} = 13.441$ with df $= 9 - 2 = 7 \Rightarrow p$-value $= .062$.

Fail to reject H_0. The conclusion that the number of fire ant hills follows a Poisson distribution appears to be supported by the data. However, we have not computed the probability of making a Type II error, so the conclusion is somewhat tenuous.

c. The fire ant hills are somewhat more clustered than randomly distributed across the pastures, although the data failed to reject the null hypothesis that the fire ant hills were randomly distributed.

10.67 a. Under the hypothesis of independence, the expected frequencies are given in the following table:

	Opinion				
Commercial	**1**	**2**	**3**	**4**	**5**
A	42	107	78	34	39
B	42	107	78	34	39
C	42	107	78	34	39

b. df $= (3 - 1)(5 - 1) = 8$

c. The cell chi-squares are given in the following table:

	Opinion				
Commercial	**1**	**2**	**3**	**4**	**5**
A	2.3810	3.7383	2.1667	4.2353	0.6410
B	2.8810	10.8037	0.0513	5.7647	21.5641
C	0.0238	1.8318	1.5513	0.1176	14.7692

$\chi^2 = \sum_{i,j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = 72.521$ with df $= 8 \Rightarrow p$-value $< .001 \Rightarrow$

Reject H_0. There is significant evidence that the commercial viewed and opinion are related.

10.78 a. Control 10%; low-dose 14%; high-dose 19%

H_0: $\pi_1 = \pi_2 = \pi_3$ versus H_a: The proportions are not all equal, where π_j is probability of a rat in group j having one or more tumors.

$E_{ij} = 100n_j/300$ and $\chi^2 = \sum_{i,j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = 3.312$ with df $= (2 - 1)(3 - 1) = 2$ and p-value $= .191$

Because the p-value is fairly large, we fail to reject H_0 and conclude there is not significant evidence of a difference in the probability of having one or more tumors for the three rat groups.

b. No, since the chi-square test failed to reject H_0.

10.81 a. The results are summarized in the following table, with $\hat{\sigma}_{\hat{\pi}} = \sqrt{(\hat{\pi})(1 - \hat{\pi})/500}$ and 95% C.I. $\hat{\pi} \pm 1.96\hat{\sigma}_{\hat{\pi}}$:

Question	$\hat{\pi}$	$\hat{\sigma}_{\hat{\pi}}$	**95% C.I.**
Did not explain?	.254	.01947	(.216, .292)
Might bother?	.916	.0124	(.892, .940)
Did not ask?	.471	.02232	(.427, .515)
Drug not changed?	.877	.0147	(.848, .906)

b. It would be important to know how the patients were selected, how the questions were phrased, the condition of the illness, and many other factors.

10.83 The combined rate for Anglo-Saxon and German: $\hat{\pi}_1 = \frac{7 + 6}{55 + 58} = .1150$

The combined rate for the other four groups: $\hat{\pi}_2 = \frac{34 + 38 + 20 + 31}{52 + 54 + 30 + 49} = .6649$

$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{(.1150)(1 - .1150)}{113} + \frac{(.6649)(1 - .6649)}{185}} = .0459$

$H_0: \pi_1 \geq \pi_2$ versus $H_a: \pi_1 < \pi_2 \quad z = \dfrac{.1150 - .6649}{.0459} = -11.98 \Rightarrow p\text{-value} < .0001 \Rightarrow$

Reject H_0 and conclude there is substantial evidence that the rate for combined group 1 is less than the rate of combined group 2.

10.85 $\bar{y} = \sum_i y_i/500 = 1.146$

After combining the last three categories, so that all $E_i > 1$ and only 1 $E_i < 5$, we obtain the following using a Poisson distribution with $\mu = 1.146$:

Mites/Leaf (k_i)	**0**	**1**	**2**	**3**	**4**	**≥ 5**
$\pi_i = P(y = k_i)$	.3179	.3643	.2088	.0797	.0228	.0065
$E_i = 500\pi_i$	158.95	182.15	104.40	39.85	11.40	3.25
n_i	233	127	57	33	30	20

$\chi^2 = \sum_i \dfrac{(n_i - E_i)^2}{E_i} = 190.57$ with df $= 6 - 1 = 5 \Rightarrow$

p-value $< .001 \Rightarrow$

Reject H_0 and conclude there is significant evidence that the data do not fit a Poisson distribution with $\mu = 1.146$.

Chapter 11: Linear Regression and Correlation

11.18 The original data and the log base 10 of recovery are given below:

Data Display

Cloud	Time	Recovery	LogRecovery
1	0	70.6	1.849
2	5	52.0	1.716
3	10	33.4	1.524
4	15	22.0	1.342
5	20	18.3	1.262
6	25	15.1	1.179
7	30	13.0	1.114
8	35	10.0	1.000
9	40	9.1	0.959
10	45	8.3	0.919
11	50	7.9	0.898
12	55	7.7	0.886
13	60	7.7	0.886

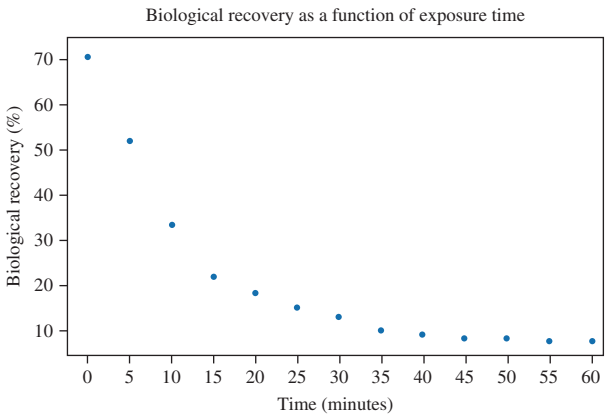
a. Scatterplot of the data is given at right.
b. Scatterplot of the data using $\log_{10}(y)$ is given at right.

11.20 $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$
Test statistic: $|t| = 9.64$
p-value $= 2P(t_4 > 9.64) < .0001 < .05 \Rightarrow$ Reject H_0 and conclude there is significant evidence that β_1 is not 0.

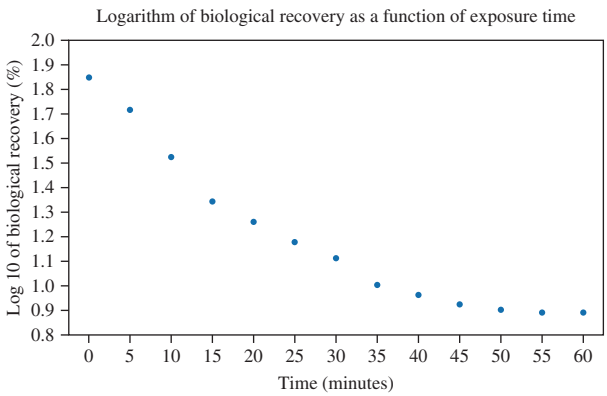
11.32 a. $\hat{y} = -1.733333 + 1.316667x$
b. The p-value for testing $H_0: \beta_1 \leq 0$ versus $H_a: \beta_1 > 0$ is p-value $= P(t_{10} \geq 6.342) < .0005 \Rightarrow$ Reject H_0 and conclude there is significant evidence that the slope β_1 is greater than 0.

11.34 a. $\hat{y} = 99.77704 + 51.9179x \Rightarrow$ When $x = 2.0, E(y) = 99.77704 + (51.9179)(2.0) = 203.613$.
b. The 95% C.I. is given in the output as (198.902, 208.323).

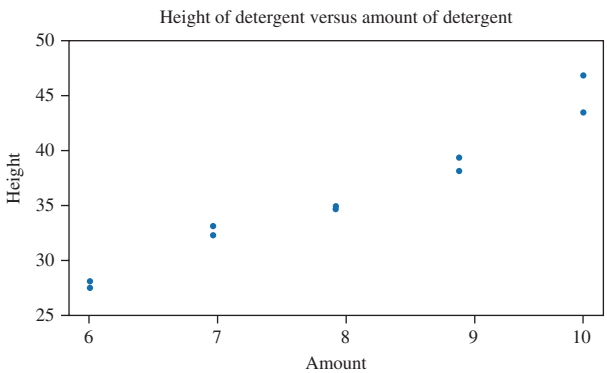
11.38 a. Scatterplot of the data is given at right.
b. $\hat{y} = 3.37 + 4.065x$
c. The residual plot indicates that higher-order terms in x may be needed in the model.



ANSWER 11.18a



ANSWER 11.18b



ANSWER 11.38a

11.74 An examination of the data in the scatterplot indicates that two of the points may possibly be outliers, since they are somewhat below the general pattern in the data. This may indicate that the data are nonnormal. Also, there appears to be an increase in the variability of the rate values as the mileage increases. This would indicate that the condition of constant variance may be violated.

11.78 a. The point is a very high influence outlier, which has distorted the slope considerably.

b. The regression line with the one point eliminated has a negative slope, $\hat{\beta}_1 = -.0015766$. This confirms the opinion of the group, which had argued that the smallest towns would have the highest per capita expenditures, with decreasing expenditures as the size of the towns increased.

11.84 a. The estimated intercept is $\hat{\beta}_0 = 53.99$. This is the estimated mean price of houses of size 0. This could be interpreted as the estimated price of land upon which there is no building. However, there were no data values with x near 0. Therefore, the estimated intercept should not be directly interpreted but just taken as a portion of an overall model.

b. A slope of 0 would indicate that the estimated mean price of houses does not increase as the size of the houses increases. That is, large houses have the same price as small houses. This is not very realistic; $t = 12.31$ with df $= 54 \Rightarrow p\text{-value} = Pr(t_{54} \geq 12.31) < .0005$. Thus, there is highly significant evidence that the slope is not 0.

c. A 95% C.I. for β_1 is $59.040 \pm (2.005)(4.794) \Rightarrow (49.428, 68.652)$.

11.88 Scatterplot of the data is given at right.
There appears to be a curvature in the plotted points, which would indicate that a straight-line model is not appropriate to model sales as a function of density.

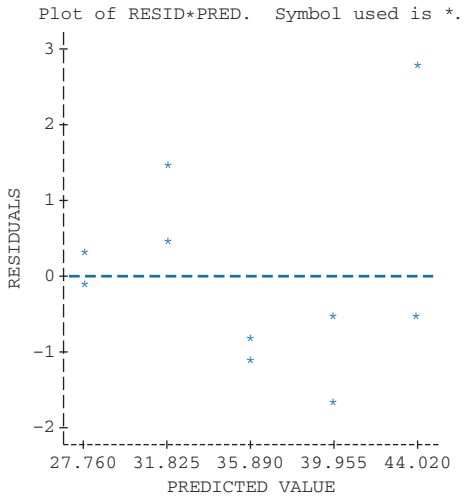
11.90 a. $\hat{y} = 47.020 + .3075x$. The estimated slope $\hat{\beta}_1 = .3075$ can be interpreted as follows: There is a .3075 increase in average durability when the concentration is increased 1 unit.

b. The coefficient of determination, $R^2 = 11.6\%$. That is, 11.6% of the variation in durability is explained by its linear relationship with concentration. Thus, a straight-line model relating durability to concentration would not yield very accurate predictions.

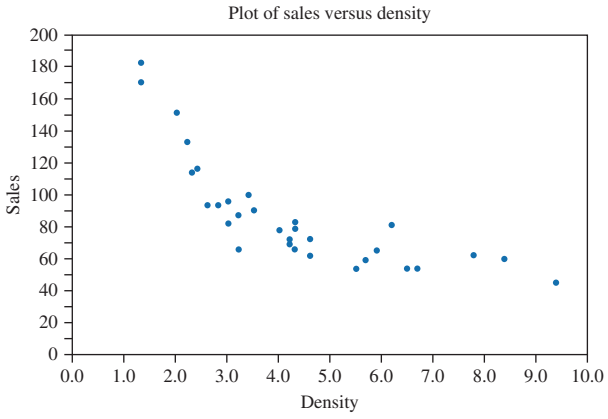
11.92 Scatterplot of the data is given at right.

a. From the scatterplot, there is a definite curvature in the relation between durability and concentration. A straight-line model would not appear to be appropriate.

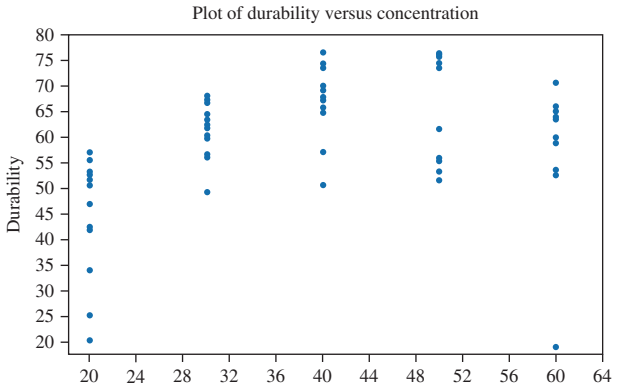
b. The coefficient of determination, R^2, measures the strength of the linear (straight-line) relation only. A straight-line model does not adequately describe the relation between durability and concentration. This is indicated by the small percentage of the variation, 11.6%, in the values of durability explained by the model containing just a linear relation with concentration. A more complex relation exists between the durability and concentration.



ANSWER 11.38c



ANSWER 11.88



ANSWER 11.92

Chapter 12: Multiple Regression and the General Linear Model

12.10 a. The logarithm of the dose levels are given here:

Dose Level (x)	2	4	8	16	32
log(x)	.693	1.386	2.079	2.773	3.466

A scatterplot of the data is given on page 1139.

b. $\hat{y} = 1.2 + 7.021 \ln(x)$

c. The model using $\ln(x)$ provides a better fit based on the scatterplot, and the residual plot appears to be a random scatter of points about the horizontal line, whereas there was a bit of curvature in the residual plot from the fit of the quadratic model.

12.12 b. No, the two independent variables, distance and population, do not appear to be severely collinear, based on the correlation ($-.24$) and the scatterplot.

c. There are two potential leverage points in the air miles direction (around 300 and 350 miles). In addition, there is one possible leverage point in the population direction; this point has a value above 200.

12.22 a. $\hat{y}$ = 7.20439 + 1.36291 METAL + .30588 TEMP + .01024 WATTS − .00277 METXTEMP

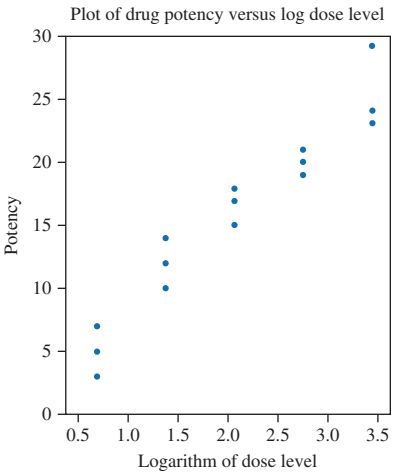
b. The results of the various t tests are given here:

H_0	H_a	T.S. t	Conclusion
$\beta_0 = 0$	$\beta_0 \neq 0$	$t = .41$	p-value = .6855 Fail to Reject H_0
$\beta_1 = 0$	$\beta_1 \neq 0$	$t = 1.47$	p-value = .1559 Fail to Reject H_0
$\beta_2 = 0$	$\beta_2 \neq 0$	$t = .19$	p-value = .8522 Fail to Reject H_0
$\beta_3 = 0$	$\beta_3 \neq 0$	$t = 2.16$	p-value = .0427 Reject H_0
$\beta_4 = 0$	$\beta_4 \neq 0$	$t = -.04$	p-value = .9717 Fail to Reject H_0

Of the four independent variables, only WATTS appears to have predictive value given the remaining three variables have already been included in the model.

c. $t_{.025,\,20} = 2.086 \Rightarrow$ 95% C.I. on β_4 is given by $-.00277 \pm (2.086)\,(.07722) \Rightarrow (-.164, .158)$.

d. VIF measures how much the standand error of a regression coefficient (β_i) is increased due to collinearity. If the value of VIF is very large, such as 10 or more, collinearity is a serious problem. The variables TEMP and METXTEMP have extremely large VIF values (250 and 246.4, respectively). An examination of the Pearson correlations reveals that the correlation between TEMP and METXTEMP is .9831—that is, nearly a perfect correlation between the two variables. One of the variables, TEMP or METXTEMP, should be removed from the model and the coefficients of the remaining variables recomputed.



ANSWER 12.10a

12.28 a. For the reduced model: R^2 is 89.53%, which is a reduction of 8.43 percentage points from the complete model's R^2 of 97.96%.

b. In the complete model, we want to test H_0: $\beta_1 = \beta_3 = 0$ versus H_a: $\beta_1 \neq 0$ and/or $\beta_3 \neq 0$.
For the reduced model, SS(Regression, Reduced) = (R^2_{Reduced})SS(Total) = (.895261)(99,379.032) = 88,970.17157.
The F statistic has the form:

$$F = \frac{[\text{SSReg., Complete} - \text{SSReg., Reduced}]/(k-g)}{\text{SS Residual, Complete}/[n-(k+1)]} = \frac{[97{,}348.339 - 88{,}970.17157]/(3-1)}{2{,}030.693/[500-4]} = 1{,}023.19$$

with df = 2, 496 $\Rightarrow$ p-value = $Pr(F_{2,496} \geq 1{,}023.19) < .0001 \Rightarrow$
Reject H_0. There is substantial evidence to conclude that $\beta_1 \neq 0$ and/or $\beta_3 \neq 0$. Based on the F test, omitting age and debt fraction from the model has substantially changed the fit of the model. Dropping one or both of these independent variables from the model will result in a decrease in the predictive value of the model.

12.32 The predicted y-value at $x = 3, w = 1, v = 6$ is $\hat{y} = 33.000$ with 95% P.I.: (21.788, 44.212). The selected values of the independent variables are at the extremes of the data used to fit the model. Therefore, the prediction is identified as being computed at "very extreme X values."

12.41 a. For testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, is p-value < .0001. Thus, we can reject H_0 and conclude there is significant evidence that the amount of additive is related to the probability of tumor development.

b. $\hat{p}\,(100) = .827$ with 95% C.I. (.669, .919)

12.47 a. $F = \dfrac{.894477/4}{(1-.894477)/(43-5)} = 80.53$ with df = 4, 38. The p-value = $Pr(F_{4,38} \geq 80.53) < .0001 \Rightarrow$
Reject H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ and conclude that at least one of the four independent variables has predictive value for loan volume.

b. Using $\alpha = .01$, none of the p-values for testing H_0: $\beta_i = 0$ versus H_a: $\beta_i \neq 0$ (.0999, .0569, .5954, and .3648, respectively) is less than .01. Thus, none of the independent variables provides substantial predictive value given the remaining three variables in the model. That is, given a model with three variables included in the model, the fourth variable does not add much when it is included.

c. The contradiction is due to the severe collinearity that is present in the four independent variables. The F test demonstrates that as a group the four independent variables provide predictive value, but because the four independent variables are highly correlated, the information concerning their relationship with the dependent variable, loan volume, is highly overlapping. Thus, it is very difficult to determine which of the independent variables are useful in predicting loan volume.

12.51 a. $\hat{y}$ = 0.8727 + 2.548 size + .220 parking + .589 income
(1.946) (1.201) (0.155) (0.178)

b. The interpretation of coefficients is given here:

Coefficient	Interpretation
$\hat{\beta}_0$ = y-intercept	The estimated average daily sales for the population of stores having 0 size, 0 parking, 0 income
$\hat{\beta}_1 = \hat{\beta}_{SIZE}$	The estimated change in average daily sales per unit change in size, for fixed values of parking and income
$\hat{\beta}_2 = \hat{\beta}_{PARKIN}$	The estimated change in average daily sales per unit change in parking, for fixed values of size and income
$\hat{\beta}_3 = \hat{\beta}_{INCOME}$	The estimated change in average daily sales per unit change in income, for fixed values of size and parking

c. $R^2 = .7912$ and $s_\varepsilon = .7724$
d. A better indicator of collinearity is the values for VIF or the R^2 values from predicting each independent variable from the remaining independent variables. Examining the correlations does not reveal any very large values. Only size and parking, with a correlation of .6565 appear to be near a value that would be of concern relative to collinearity.

12.53 a. $\hat{y} = 102.708 - .833$ PROTEIN $- 4.000$ ANTIBIO $- 1.375$ SUPPLEM
b. $s_\varepsilon = 1.70956$
c. $R^2 = 90.07\%$
d. There is no collinearity problem in the data set. The correlations between the pairs of independent variables are 0 for each pair, and the VIF values are all equal to 1.0. This total lack of collinearity is due to the fact that the independent variables are perfectly balanced. Each combination of PROTEIN and ANTIBIO values appears exactly three times in the data set. Each combination of PROTEIN and SUPPLEM occur twice, and so on.

12.55 a. $\hat{y} = 89.8333 - .83333$ PROTEIN
b. $R^2 = .5057$
c. In the complete model, we want to test
$H_0: \beta_2 = \beta_3 = 0$ versus H_a: At least one of $\beta_2, \beta_3 \neq 0$.
The F statistic has the form:

$$F = \frac{[371.083 - 208.333]/(3 - 1)}{40.9166/[18 - 4]} = 27.84$$

with df $= 2, 14 \Rightarrow p$-value $= Pr(F_{2,14} \geq 27.84) < .0001 \Rightarrow$ Reject H_0.
There is substantial evidence to conclude that at least one of $\beta_2, \beta_3 \neq 0$. Based on the F test, omitting x_2 and/or x_3 from the model would substantially change the fit of the model. Dropping ANTIBIO and/or SUPPLEM from the model may result in a large decrease in the predictive value of the model.

12.57 a. $R^2 = .3844 = 38.44\%$
b. R^2 has decreased dramatically to $.0358 = 3.58\%$.
c. In the complete model, we want to test
$H_0: \beta_2 = \beta_3 = 0$ versus H_a: At least one of $\beta_2, \beta_3 \neq 0$.
The F statistic has the form:

$$F = \frac{[39.31706 - 3.66167]/(3 - 1)}{62.95698/[67 - 4]} = 17.93$$

with df $= 2, 63 \Rightarrow p$-value $= Pr(F_{2,63} \geq 17.93) < .0001 \Rightarrow$ Reject H_0.
There is substantial evidence to conclude that at least one of $\beta_2, \beta_3 \neq 0$. Based on the F test, omitting MARGIN and/or IPCOST from the model would substantially change the fit of the model. Dropping MARGIN and IPCOST from the model will result in a large decrease in the predictive value of the model.

12.61 When NUMEMP $= 500$, SIZE $= 2.5$, PERSCOSTS $= 55$, $\hat{y} = 69.7627\%$, and a 95% P.I. for y is $(58.1829\%, 81.3424\%)$. The value 88.9% falls outside the P.I. and hence would appear to be somewhat unreasonable in this situation.

Chapter 13: Further Regression Topics

13.21 a. The estimated coefficient associated with promotion is -19.960. This indicates that for fixed values of price and category, the average value of sales is estimated to be reduced by 19.960 if a competing brand is having a promotion; otherwise, the average value of sales does not change.
b. One would suspect that a promotion by a truly competing brand would result in a decrease in sales. The model predicts this result, since the estimated coefficient is negative.
c. The t statistic for testing whether the promotion coefficient if different from 0 has p-value $< .0001$. Thus, there is significant evidence that the promotion coefficient differs from 0.

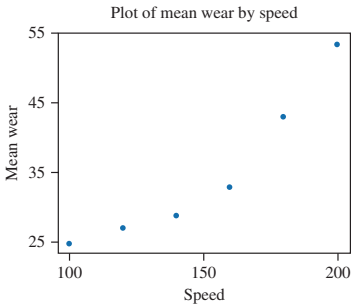
13.23 When promotions are offered by a competing brand, PROMOTION $= 1$, the model becomes:
$\hat{y} = 26.807 + 90.233$ PRICE $+ .134$ CATEGORY $+ 287.609(1) - 142.433$ (PRICE)(1) $- .024$ (CATEGORY)(1)
$\hat{y} = 314.416 - 52.200$ PRICE $+ .110$ CATEGORY
When promotions are not offered by a competiting brand, PROMOTION $= 0$, the model becomes:
$\hat{y} = 26.807 + 90.233$ PRICE $+ .134$ CATEGORY $+ 287.609(0) - 142.433$ (PRICE)(0) $- .024$ (CATEGORY)(0)
$\hat{y} = 26.807 + 90.233$ PRICE $+ .134$ CATEGORY
The models for predicting sales have considerably different intercepts depending on whether or not there is a promotion for a competing brand. The partial slopes for PRICE for the two models have different signs and very different magnitudes. The change in sign is of interest. It demonstrates that when there is a promotion for a competing brand, if the price is increased, sales drop considerably, whereas if there is not a promotion for a competing brand, a price increase does not result in a decrease in sales.

13.31 a. $\hat{y} = -2.704 + .517$ RATE5 $+ 1.450$ UNEMPLOY $+ .0353$ RT5*UNEP
The fitted model has $R^2 = 92.67\%$, and the three residual plots do not indicate any major pattern; thus, the model appears to fit quite well.
b. A check of model conditions:
1. Zero expectation: The model appears to not need any higher-order terms.
2. Constant variance: From the residuals versus predicted values, there does not appear to be an indication of unequal variation.
3. Normality: The boxplot appears slightly skewed to the right, but there are no outliers. There is a slight indication of nonnormality in the normal probability plots. Neither of these indications appears to require a transformation of the data.
4. The Durbin–Watson statistic equals 2.403, which would indicate a mild negative serial correlation, but because it is less than 2.5, a differencing of the data is probably unnecessary.

13.33 The residual plot indicates that the model is underestimating y for small values of $\hat{y}$ and overestimating y for large values of $\hat{y}$. Thus, additional terms may be needed in the model. Since the data are quarterly earnings, there is the possibility of serial correlation. A plot of the residuals versus time would be recommended.

13.37 b. Linear model: $\hat{y} = 8.667 + .575\ \text{DOSE}$
Quadratic model: $\hat{y} = 4.484 + 1.506\ \text{DOSE} - .0270\ (\text{DOSE})^2$

c. The quadratic model appears to be more appropriate: It has a larger R^2 (88.15% versus 77.30%) and a smaller MS(Error) (7.548 versus 13.345); the term $DOSE^2$ has p-value = .0062, which indicates that the quadratic term significantly improves the fit in comparison to the linear model; and the residuals are somewhat smaller in the quadratic model with a less apparent pattern when compared to the residuals from the linear model.

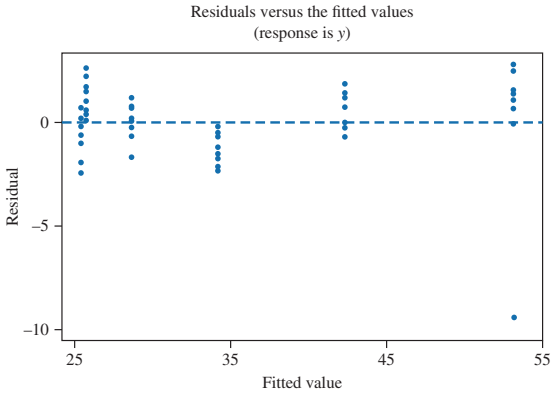


ANSWER 13.41a

13.41 a. A scatterplot of the data is given at right.
It would appear that a quadratic model in machine speed is needed.

b. The estimated regression equation is
$\hat{y} = 63.139 - .70507x_1 + .0032768x_1^2$

c. A residual plot for the fitted model is given at right.
It would appear that the model is not an adequate representation of the variation in wear, since at some machine speeds all the residuals are positive and at other machine speeds all the residuals are negative. Although the model overall is providing an excellent fit to the data, this pattern would indicate that further modeling is needed. For example, there may be other independent variables besides machine speed that may affect wear.



ANSWER 13.41c

13.43 The first fitted regression equation is

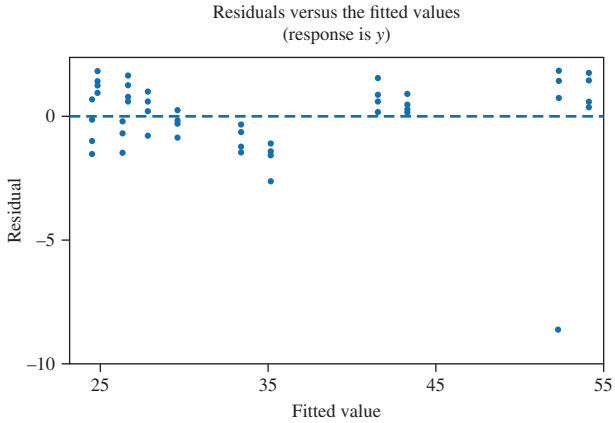
$$\hat{y} = 60.477 - .705x_1 + .00328x_1^2 + 8.875x_2$$

The second fitted regression equation is

$$\hat{y} = 42.28 - .421x_1 + .00224x_1^2 + 69.54x_2 - .949x_1x_2 + .00345x_1^2x_2$$

These two models provide only marginal improvement over the quadratic model in just x_1. However, the pattern in the residual plot noted from the quadratic model in x_1 is not as noticeable in the residual plots from these two models.

A residual plot of the first fitted model is given at right.



ANSWER 13.43

13.45 There is no indication of the plot of height by amount of a quadratic curvature. Hence, the second-order terms in amount are probably unnecessary.

13.49 a. The fitted model is $\hat{y} = 44.182 - .494x + .00143x^2$.

b. From the output, $F = \dfrac{36.7/3}{91.8/9} = 1.20 \Rightarrow p\text{-value} = .364$.

Thus, there is not significant evidence of lack of fit of the model; higher-order terms in temperature (x) are not needed to adequately fit the data.

c. There are no obvious patterns in the residual plot.

13.51 The calculations for the test of lack of fit are given here:

x (Dose Level)	$\bar{y}_i$	$\Sigma_j (y_{ij} - \bar{y}_i)^2$	$n_i - 1$
2	5	8	2
4	12	8	2
8	16.667	4.667	2
16	20	2	2
32	25.333	20.667	2
Total		43.334	10

$\text{SSP}_{\text{exp}} = 43.334, \text{df}_{\text{exp}} = 10$

From the output from Exercise 13.37, SS(Residual) = 90.579 and $\text{df}_{\text{Residual}} = 12$.

The $\text{SS}_{\text{Lack}} = 90.579 - 43.334 = 47.245$ and $\text{df}_{\text{Lack}} = 12 - 10 = 2$.

$$F = \frac{47.245/2}{43.334/10} = 5.45 \quad \text{df} = 2, 10 \Rightarrow p\text{-value} = .0251$$

There is significant evidence of lack of fit of the quadratic model. Hence, higher-order terms in dose level, such as x^3 and x^4 may be required to improve the fit of the model.

13.63 **a.** The question is a test of $H_0: \beta_1 = \beta_2 = 0$ versus $H_a: \beta_1 \neq 0$ and/or $\beta_2 \neq 0$.

From the output, $F = \dfrac{\text{MS(Model)}}{\text{MS(Error)}} = 15.987$ with p-value $< .0001 < .05 \Rightarrow$

Reject H_0 and conclude there is significant evidence that ROOMS and SQUARE FEET, taken together, contain information about PRICE.

b. Test $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$;

$t = .717$ with p-value $= .4822 > .05 \Rightarrow$

Fail to reject H_0 and conclude there is not significant evidence that the coefficient of ROOMS is different from 0.

c. Test $H_0: \beta_2 = 0$ versus $H_a: \beta_2 \neq 0$;

$t = 1.468$ with p-value $= .1585 > .05 \Rightarrow$

Fail to reject H_0 and conclude there is not significant evidence that the coefficient of SQUARE FEET is different from 0.

13.65 The F test of the overall model is 4.42 with p-value $= .0041$.
The indicator variable $RC3$ measures the difference in risk of infection between hospitals in the south and west, holding all other variables constant. The coefficient of $RC3$ is β_7, and we want to test $H_0: \beta_7 \leq .5\%$ versus $H_a: \beta_7 > .5\%$. The test statistic is

$$t = \frac{\beta_7 - .5}{SE(\beta_7)} = \frac{.7024 - .5}{.8896} = .23 \text{ with df} = 20$$

p-value $= Pr(t_{20} > .23) = .4102 \Rightarrow$

Fail to reject H_0; there is not significant evidence that the infection rate in the south is at least .5% higher than in the west.

13.67 The following model is selected:
$y = \beta_0 + \beta_1 \text{STAY} + \beta_3 \text{INS} + \varepsilon$

The R^2 for this model is .5578 versus .6072 for the seven-variable model.

The MS(Error) for this model is 28.765 versus 25.546 for the seven-variable model.

A test of H_0: Two-variable model versus H_a: Seven-variable model is given by testing the following parameters in the seven-variable model:

$H_0: \beta_2 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ versus H_a: At least one of $\beta_2, \beta_4, \beta_5, \beta_6, \beta_7 \neq 0$;

$$F = \frac{(39.49805177 - 36.27961297)/5}{25.54623394/20} = .50 \text{ with df} = 5, 20 \Rightarrow p\text{-value} = Pr(F_{5,20} > .50) = .7726 \Rightarrow$$

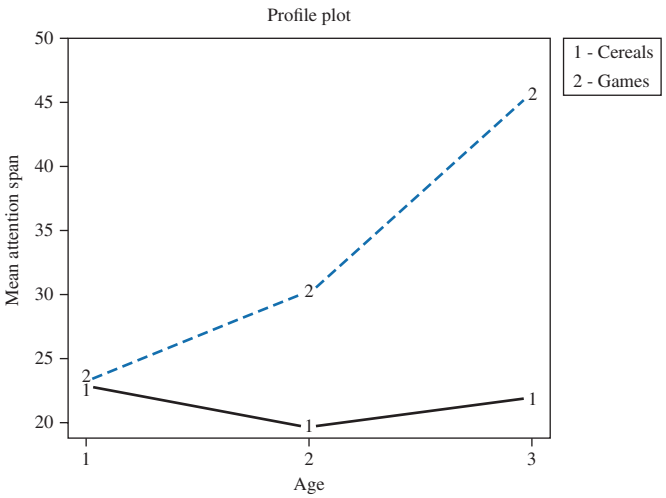
Fail to reject H_0; there is not significant evidence that any of the five parameters is not 0. Thus, there is not significant evidence of a difference between the two-variables and seven-variables models.

Based on the above test, the marginal difference in R^2, and MS(Error), the model with fewer variables is the more desirable model.

Chapter 14: Analysis of Variance for Completely Randomized Designs

14.9 **a.** A profile plot of the data is given at right.
The profile plot indicates an increasing effect of product type as age increases.

b. The p-value for the interaction term is .013. There is significant evidence of an interaction between the factors age and product type. Thus, the amount of difference in mean attention spans of children between breakfast cereals and video games would vary across the three age groups. From the profile plots, the estimated mean attention span for video games is larger than for breakfast cereals, with the size of the difference becoming larger as age increases.



ANSWER 14.9a

14.17 The necessary parameters are $t = 8, D = 20, \alpha = .05$,

$$\sigma = 9 \Rightarrow \varphi = \sqrt{\frac{r(20)^2}{(2)(6)(9)^2}} = .5556\sqrt{r}$$

Determine r so that power is .80. Select values for r; compute $\nu_1 = t - 1 = 8 - 1 = 7$, $\nu_2 = t(r-1) = 8(r-1)$, and $\phi = 5556\sqrt{r}$; then use Table 14 with $\alpha = .05$ and $t = 8$ to determine power:

r	ν_2	ϕ	Power
5	32	1.24	.61
6	40	1.36	.73
7	48	1.47	.82

Thus, it would take seven reps to obtain a power of at least .80.

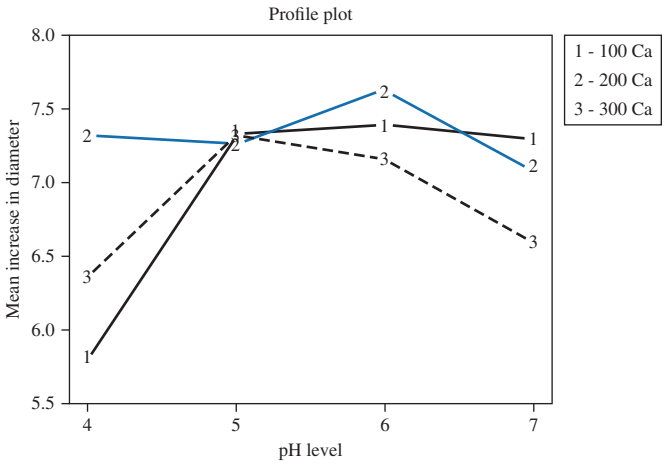
14.21 a. The test for an interaction has $F = 11.34$ with df $= 9, 16$ which yields a p-value $= 0.0001$. This implies there is significant evidence of an interaction between Cu rate and Mn rate on soybean yield.

b. Mn $= 110$

c. Cu $= 7$

d. (Cu, Mn) $= (7, 110)$

14.23 a. The profile plot is given at right.
There appears to be an interaction between Ca rate and pH with respect to the increase in trunk diameters. At low pH value, a 200 level of Ca yields the largest increase, whereas at high pH value, a 100 level of Ca yields the largest increase in trunk diameter.



ANSWER 14.23a

b. A model for this experiment is given here:

$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}; i = 1, 2, 3, 4; j = 1, 2, 3; k = 1, 2, 3$

where y_{ijk} is the increase in trunk diameter of the kth tree in soil having the ith pH level using the jth Ca rate,

τ_i is the effect of the ith pH level on diameter increase,

β_j is the effect of the jth Ca rate on diameter increase, and

$\tau\beta_{ij}$ is the interaction effect of the ith pH level and jth Ca rate on diameter increase.

c. This is a completely randomized 4×3 factorial experiment with factor A: pH level and factor B: Ca rate. There are three complete replications of the experiment. The AOV table is given here:

Source	df	SS	MS	F	p-value
pH	3	4.461	1.487	21.94	.0001
Ca	2	1.467	.734	10.82	.0004
Interaction	6	3.255	.543	8.00	.0001
Error	24	1.627	.0678		
Total	35	10.810			

14.25 a. Using Tukey's W procedure with $\alpha = .05, s_\varepsilon^2 = \text{MSE} = .0678, q_\alpha(t, \text{df}_{\text{error}}) = q_{.05}(3, 24) = 3.53 \Rightarrow W = (3.53)\sqrt{\frac{.0678}{3}} = .53 \Rightarrow$

		Ca Rate		
		100	200	300
pH = 4	Mean	5.80	7.33	6.37
	Grouping	a	c	b
pH = 5	Mean	7.33	7.27	7.33
	Grouping	a	a	a
pH = 6	Mean	7.40	7.63	7.17
	Grouping	a	a	a
pH = 7	Mean	7.30	7.10	6.60
	Grouping	b	ab	a

b. From the above table, we observe that at pH $= 5, 6$ there is not significant evidence of a difference in mean increases in diameter between the three levels of Ca. However, at pH $= 4, 7$ there is significant evidence of a difference, with Ca $= 200$ yielding the largest increase at pH $= 4$ and Ca $= 100$ or 200 yielding the largest increase at pH $= 7$. This illustrates the interaction between Ca and pH; i.e., the size of differences in the means across the levels of Ca depends on the level of pH.

14.27 a. The design is a completely randomized 3×9 factorial experiment with five replications; factor A is level of severity and factor B is type of medication.

b. A model for this experiment is given here:

$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}; \quad i = 1, 2, 3; \quad j = 1, \cdots, 9; \quad k = 1, 2, 3, 4, 5$

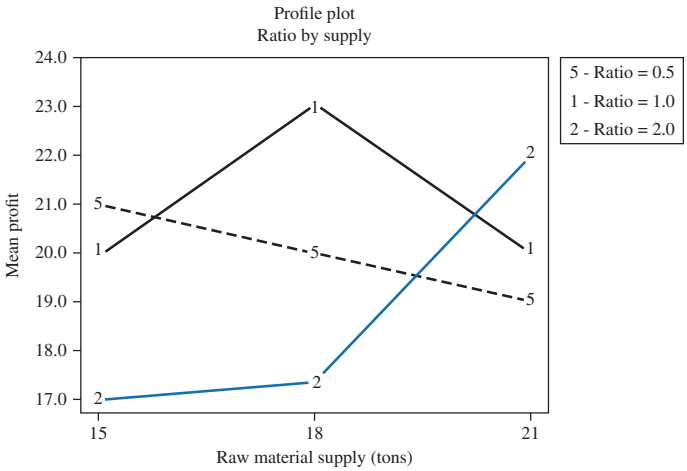
where y_{ijk} is the temperature of the kth patient having the ith severity level using the jth medication,

τ_i is the effect of the ith severity level on temperatures,

β_j is the effect of the jth medication on temperature, and

$\tau\beta_{ij}$ is the interaction effect of the ith severity level and jth medication on temperature.

14.35 **a.** The test for an interaction yields p-value = .0255. There is significant evidence that an interaction exists between ratio and supply in regard to the mean profit. The profile plot on the right displays the interaction.



ANSWER 14.35a

Chapter 15: Analysis of Variance for Blocked Designs

15.7 The model conditions appear to be satisfied:

The normal probability plots and boxplots of the residuals do not indicate nonnormality.

Plot of residuals versus estimated mean does not indicate nonconstant variance.

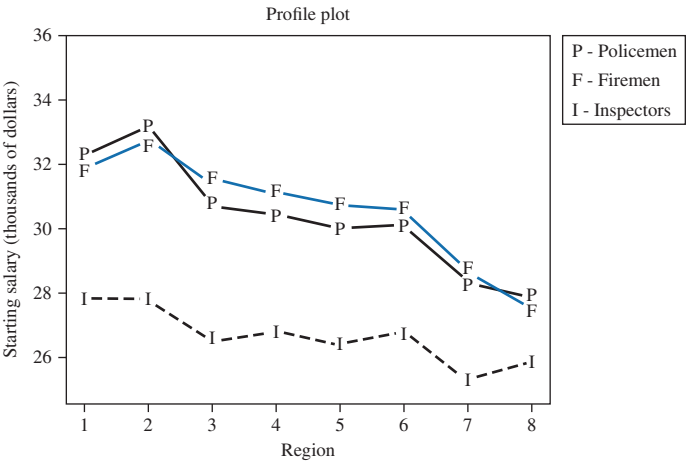
Interaction plot indicates a potential interaction between subjects and type of music, but the indications are fairly weak.

15.11 **a.** The boxplot and normal probability plot do not indicate a deviation from a normal distribution for the residuals.

The plot of residuals versus estimated means does not indicate a deviation from the constant variance condition.

Based on these plots, there is no indication of any deviations from the model conditions.

15.29 **a.** A profile plot of the data is given at right.



ANSWER 15.29a

Based on the profile plot, the additive model appears to be appropriate because the three lines are relatively parallel. Note further that the plotted points are means of a single observation and hence may be quite variable in their estimation of the population means μ_{ij}. Thus, exact parallelism is not required in the profile plots to ensure the validity of the additive model.

It would not be possible to test for an interaction between region and job type because there is only one observation per region–job type combination.

b. $$\text{RE(RCB,CR)} = \frac{(b-1)\text{MSB} + b(t-1)\text{MSE}}{(bt-1)\text{MSE}} = \frac{(8-1)(6.089) + (8)(3-1)(.422)}{((8)(3)-1)(.422)} = 5.09 \Rightarrow$$

It would take 5.09 times as many observations (approximately 41) per treatment in a completely randomized design to achieve the same level of precision in estimating the treatment means as was accomplished in the randomized complete block design.

c. Other possible important factors may be average salaries of all government employees in the region, education requirements for the position, and so on.

15.33 **a.** Randomized complete block design with the five specimens of fabrics serving as the blocks and the three dyes being the treatments.

b. The test for the differences in mean quantities of the three dyes has p-value = .0100. Thus, there is significant evidence of a difference in the mean quantities of the three dyes.

Using Tukey's W procedure with $\alpha = .05, s_\varepsilon^2 = \text{MSE} = 34.367, q_\alpha(t, \text{df}_{\text{error}}) = q_{.05}(3, 8) = 4.04 \Rightarrow$

$$W = (4.04)\sqrt{\frac{34.367}{5}} = 10.59 \Rightarrow$$

	Dye		
	A	**B**	**C**
Mean	77.40	84.60	92.80
Grouping	a	ab	b

c. $t = 3, b = 5 \Rightarrow \text{RE(RCB,CR)} = \dfrac{(5-1)(23.567) + (5)(3-1)(34.367)}{((5)(3)-1)(34.367)} = .91 \Rightarrow$

It would take .91 times as many observations (approximately 5) per treatment in a completely randomized design to achieve the same level of precision in estimating the treatment means as was accomplished in the randomized complete block design. Since RE was slightly less than 1, we would conclude that the blocking was not effective.

15.35 **a.** Latin square design with blocking variables farm and plot. The treatment is the five types of fertilizers.

b. There is significant evidence (p-value $< .0001$) the mean yields are different for the five fertilizers.

Chapter 16: The Analysis of Covariance

16.15 **a.** Randomized complete block design with the three antidepressants as treatments, the age–gender combinations as six blocks, and the pretreatment rating serving as a covariate.

b. $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i}x_{2i} + \beta_5 x_{1i}x_{3i} + \beta_6 x_{4i} + \beta_7 x_{5i} + \beta_8 x_{6i} + \beta_9 x_{7i} + \beta_{10} x_{8i} + \varepsilon_i$ for $i = 1, \ldots, 16$

x_1 = covariate

$$x_2 = \begin{cases} 1 & \text{if antidepressant B} \\ 0 & \text{if otherwise} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{if antidepressant C} \\ 0 & \text{if otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if observation in block 2} \\ 0 & \text{if otherwise} \end{cases} \qquad x_5 = \begin{cases} 1 & \text{if observation in block 3} \\ 0 & \text{if otherwise} \end{cases}$$

$$x_6 = \begin{cases} 1 & \text{if observation in block 4} \\ 0 & \text{if otherwise} \end{cases} \qquad x_7 = \begin{cases} 1 & \text{if observation in block 5} \\ 0 & \text{if otherwise} \end{cases}$$

$$x_8 = \begin{cases} 1 & \text{if observation in block 6} \\ 0 & \text{if otherwise} \end{cases}$$

16.19 **a.** Test for parallelism of the four treatment lines:

$$F = \frac{(3{,}316.8281 - 3{,}180.7299)/(75-72)}{3{,}180.7299/72} = 1.03, \text{ with df} = 3, 72 \Rightarrow$$

p-value $= Pr(F_{3,72} \geq 1.03) = .385 \Rightarrow$

There is not significant evidence that the lines are not parallel.

b. Test for difference in adjusted treatment means:

$$F = \frac{(8{,}724.7852 - 3{,}316.8281)/(78-75)}{3{,}316.8281} = 40.76, \text{ with df} = 3, 75 \Rightarrow$$

p-value $= Pr(F_{3,75} \geq 40.76) < .0001 \Rightarrow$

There is significant evidence that the adjusted mean ratings are different for the four socioeconomic classes.

c. $\hat{\mu}_{adj,1} = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 \bar{x}.. = (37.197 - 22.490) + (.27472)(28.95) = 22.66$

$\hat{\mu}_{adj,2} = (\hat{\beta}_0 + \hat{\beta}_3) + \hat{\beta}_1 \bar{x}.. = (37.197 - 15.951) + (.27472)(28.95) = 29.20$

$\hat{\mu}_{adj,3} = (\hat{\beta}_0 + \hat{\beta}_4) + \hat{\beta}_1 \bar{x}.. = (37.197 - 14.784) + (.27472)(28.95) = 30.37$

$\hat{\mu}_{adj,4} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}.. = 37.197 + (.27472)(28.95) = 45.15$

$$\text{SE}(\hat{\mu}_{adf,1}) = \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{(\bar{x}_{1.} - \bar{x}..)^2}{E_{xx}}\right)} = \sqrt{(44.2244)\left(\frac{1}{20} + \frac{(28.95 - 28.95)^2}{9135.8}\right)} = 1.4870$$

$$\text{SE}(\hat{\mu}_{adf,2}) = \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{(\bar{x}_{2.} - \bar{x}..)^2}{E_{xx}}\right)} = \sqrt{(44.2244)\left(\frac{1}{20} + \frac{(28.70 - 28.95)^2}{9135.8}\right)} = 1.4871$$

$$\text{SE}(\hat{\mu}_{adf,3}) = \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{(\bar{x}_{3.} - \bar{x}..)^2}{E_{xx}}\right)} = \sqrt{(44.2244)\left(\frac{1}{20} + \frac{(28.60 - 28.95)^2}{9135.8}\right)} = 1.4872$$

$$\text{SE}(\hat{\mu}_{adf,4}) = \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{(\bar{x}_{4.} - \bar{x}..)^2}{E_{xx}}\right)} = \sqrt{(44.2244)\left(\frac{1}{20} + \frac{(29.55 - 28.95)^2}{9135.8}\right)} = 1.4876$$

$t_{1-(.05)v(2)(4),\,75} = t_{.00625,\,75} = 2.559$

95% C.I.s for the mean adjusted verbalization scores:

Socioeconomic class 1: $22.66 \pm (2.559)(1.4870) \Rightarrow (18.9, 26.5)$

Socioeconomic class 2: $29.20 \pm (2.559)(1.4871) \Rightarrow (25.4, 33.0)$

Socioeconomic class 3: $30.37 \pm (2.559)(1.4872) \Rightarrow (26.6, 34.2)$

Socioeconomic class 4: $45.15 \pm (2.559)(1.4876) \Rightarrow (41.3, 49.0)$

The four confidence intervals indicate that socioeconomic classes 1, 2, and 3 had similar adjusted mean verbalization scores, but socioeconomic class 4 appears to have considerably higher scores than the other three classes.

Chapter 17: Analysis of Variance for Some Fixed-, Random-, and Mixed-Effects Models

17.17 **a.** The mixed-effects model is most appropriate. The researcher would be concerned about specific chemicals, not a population of chemicals. He would want to determine which of the four chemicals is most effective in controlling fire ants.

b. A fixed-effects model would be appropriate if the researcher was interested only in a set of specific locations, such as those with specific environmental conditions, or different levels of human activity or specific soil conditions. The fixed-effects model would have both the levels of chemicals and the levels of locations used in the experiment as the only levels of interest. The levels used in the experiment were not randomly selected from a population of levels.

17.23 **a.** A test for the equality of the treatment means in the fixed-effects model is

H_0: $\tau_1 = \cdots = \tau_t = 0$ versus H_a: At least one τ_i is not 0.

In the fixed-effects model, we are testing the difference in the means for the t treatments used in the experiment.

b. A test concerning the variability in the population of means in the random-effects model is

H_0: $\sigma_\tau^2 = 0$ versus H_a: $\sigma_\tau^2 > 0$.

In the random- effects model, we are testing the difference in a population of means from which the t treatments used in the experiment were randomly selected, and not just the means used in the study.

17.25 **a.** This is two reps of a completely randomized mixed model with

Factor A: Temperature is fixed with five levels

Factor B: Pane design is random with five levels

The AOV table is given here:

Source	df	SS	MS	EMS	*F*	*p*-value
Temp	4	39.7788	9.9447	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2 + 10\theta_\tau$	14.50	.0001
Panes	4	7.3228	1.8307	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2 + 10\sigma_\beta^2$	2.67	.0703
Interaction	16	10.9712	.6857	$\sigma_\varepsilon^2 + 2\sigma_{\tau\beta}^2$	2.97	.0072
Error	25	5.7800	.2312	σ_ε^2		
Total	49	63.8528				

b. The interaction between temperature and pane design is significant (p-value = .0072), the main effect of temperature is significant (p-value < .0001), but the main effect of pane design is not significant (p-value = .0703).

c. In Exercise 14.31, all three terms were also significant at essentially the same p-values. Another difference is that in this case the inferences made concern the population of pane designs and not just the five designs used in the study.

d. If there is a very large number of commercial thermal pane designs available, then it would be reasonable to randomly select a few for comparison in the study. If the only pane designs available are the five used in the study, then the fixed-effects model would be the appropriate model.

17.31 **a.** This is a nested design with samples nested within batches.

b. A model for this situation is:

$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \varepsilon_{ijk}$

where y_{ijk} is the hardness of the kth tablet from sample j selected from batch i,

μ is the overall mean hardness,

τ_i is the random batch effect, iid $N(0, \sigma_\tau^2)$,

$\beta_{j(i)}$ is the random sample within batch effect, iid $N(0, \sigma_{\beta(\tau)}^2)$,

ε_{ijk} is the random effect due to all other factors, iid $N(0, \sigma_\varepsilon^2)$, and

$\tau_i, \beta_{j(i)}$, and ε_{ijk} are all independent.

c. The AOV table is given here:

Source	df	SS	MS	*F*	*p*-value
Batch	2	9,095.5238	4,547.7619	101.635	.0001
Sample	6	268.4762	44.7460	1.533	.1851
Error	54	1,576.0000	29.1852		
Total	62	10,940.0000			

d. There is significant evidence (p-value < .0001) that the batches produced different mean hardness values. There does not appear to be a significant (p-value = .1851) variation in the samples within the batches.

The variance components are given here:

Source	Var Component	% of Total
Batch	214.429	87.22
Sample	2.223	0.90
Error	29.185	11.87
Total	245.837	

The major source of variation in hardness of the tablets is due to the batch-to-batch variation.

Chapter 18: Split-Plot, Repeated Measures, and Crossover Designs

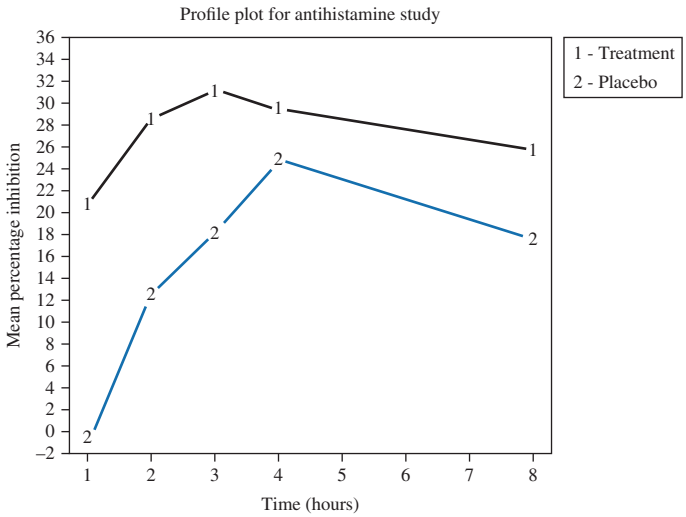
18.7 b. There appears to be an increase in the mean water loss as the level of saturation deficit increases.

18.9 a. The mean and standard deviation of percentage inhibition by treatment and time are given here:

	Time				
Treatment (Means)	**1**	**2**	**3**	**4**	**8**
Antihistamine	20.70	28.57	31.24	29.44	25.63
Placebo	−0.76	12.55	18.23	24.79	17.57
Treatment (St. Dev.)	**1**	**2**	**3**	**4**	**8**
Antihistamine	23.98	12.00	14.30	12.65	14.26
Placebo	12.26	10.43	10.83	6.91	7.83

The antihistamine-treated patients uniformly, across all five hours, have larger mean percentage inhibitions than the placebo-treated patients. The pattern for the standard deviations is similar, with somewhat higher values during the first hour after treatment.

b. A profile plot of the skin sensitivity data is given here:



Yes, the antihistamine-treated patients appear to have higher mean percentage inhibitions than the placebo-treated patients with the size of the difference between the placebo and antihistamine patients fairly consistent across the five hours of measurements.

18.19 Based on the results in the AOV table, the conclusions based on the profile plot are confirmed. There is a significant period effect (p-value $< .0001$), the effect due to formulations is not significant, (p-value $= .733$), and there is not an effect due to sequence (p-value $= .071$).

Chapter 19: Analysis of Variance for Some Unbalanced Designs

19.21 a. $SST_{adj} = SSE_{red.1} - SSE_{complete} = 100.21 - 17.91 = 82.3$, with df $= 8 - 5 = 3$

$SSR_{adj} = SSE_{red.2} - SSE_{complete} = 25.40 - 17.91 = 7.49$, with df $= 8 - 5 = 3$

$SSC_{adj} = SSE_{red.3} - SSE_{complete} = 713.00 - 17.91 = 695.10$, with df $= 8 - 5 = 3$ with df $= 18 - 3 - 11 = 4$

Summarize these values in an AOV table:

Source	**df**	**SS**	**MS**	***F***	***p*-value**
Blend (corrected)	3	82.30	27.43	7.66	.0257
Driver (corrected)	3	7.49	*	*	*
Model (corrected)	3	695.10	*	*	*
Error	5	17.91	3.58	*	*
Total	14	806.58	*	*	*

19.23 **c.** The following table contains the intermediate calculations needed to obtain the sum of squares for the treatment:

Block	Block Total	Block Mean
1	106	35.33
2	125	41.667
3	115	38.333
4	115	38.333
5	107	35.667
6	157	52.333
7	142	47.333
8	116	38.667
9	154	51.333
10	127	42.333

Treatment	A	B	C	D	E	F	Total
y_i	211	175	284	172	171	251	1,264
B_i	642	580	695	595	640	640	
$3y_{i.} - B_i$	−9	−55	157	−79	−127	113	0
$(3y_{i.} - B_i)^2$	81	3,025	24,649	6,241	16,129	12,769	62,894

$\bar{y}.. = 1{,}264/30 = 42.133$

$\text{TSS} = \sum_{ij} (y_{ij} - 42.133)^2 = 3{,}235.467$

$\text{SSB} = k\sum_j (\bar{y}_{.j} - \bar{y}..)^2 = 3\sum_j (\bar{y}_{.j} - 42.133)^2 = 1{,}034.80$

$$\text{SST}_{\text{adj}} = \frac{t-1}{nk(k-1)}\sum_i (ky_{i.} - B_{(i)})^2 = \frac{6-1}{(30)(3)(3-1)}(62{,}894) = 1{,}747.056$$

$\text{SSE} = \text{TSS} - \text{SST}_{\text{adj}} - \text{SSB} = 3{,}235.467 - 1{,}747.056 - 1{,}034.8 = 453.611$

Summarizing in an AOV table:

Source	df	SS	MS	*F*	*p*-value
Treatment (ADJ)	5	1,747.056	349.411	11.55	.0001
Block	9	1034.8	*	*	*
Error	15	4,53.6111	30.241	*	*
Total	29	3,235.467	*	*	*

Because the *p*-value $< .0001$, we conclude that there is significant evidence that the six antihistamines have different mean responses.

19.24 The adjusted treatment means are obtained from the equation:

$$\hat{\mu}_i = \bar{y}.. + \frac{ky_{i.} - B_{(i)}}{t\lambda} = 42.133 + \frac{3y_{i.} - B_{(i)}}{(6)(2)}$$

$\text{MSE} = 30.241 \quad \text{df}_{\text{Error}} = 15 \quad t_{.025,15} = 2.131$

$$W = \frac{q_{.025}(6,15)}{\sqrt{2}}\sqrt{\frac{2k\text{MSE}}{t\lambda}} = 12.64$$

The calculations are summarized in the following table:

Treatment	A	B	C	D	E	F
$\bar{y}_{i.}$	42.2	35	56.8	34.4	34.2	50.2
$3y_{i.} - B_i$	−9	−55	157	−79	−127	113
$\hat{\mu}_i$	41.38	37.55	55.22	35.55	31.55	51.54

The groupings based on LSD are given here:

Treatment	E	D	B	A	F	C
$\hat{\mu}_i$	31.55	35.55	37.55	41.38	51.54	55.22
Groups	a	a	a	ab	bc	c

The treatments with common letters are not significantly different. Thus, the significantly different pairs of treatments are (E,F), (E,C), (D,F), (D,C), (B,F), (B,C), (A,C).

19.25 **a.** The experiment consists of the same three chains observed in four different geographical areas. In each area, we obtain the weekly sales volume during two different weeks for each of the three chains. This is a randomized complete block experiment with blocks (weeks) and treatments consisting of a 4×3 factorial structure with factors area and chain. The model for this situation is:

$y_{ijk} = \mu + \gamma_k + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$

where y_{ijk} is the sales volume during week k at chain i in area j,

γ_k is the effect of week k,

τ_i is the effect of chain i,

β_j is the effect of area j,

$\tau\beta_{ij}$ is the interaction effect of chain i in area j, and

ε_{ijk} is the random effect of all other factors.

b. The study would then simply be a single replication of a complete randomized design with treatments consisting of a 4×3 factorial structure with factors area and chain. Since there is only a single replication, the interaction term cannot be estimated or tested. The model would reduce to:

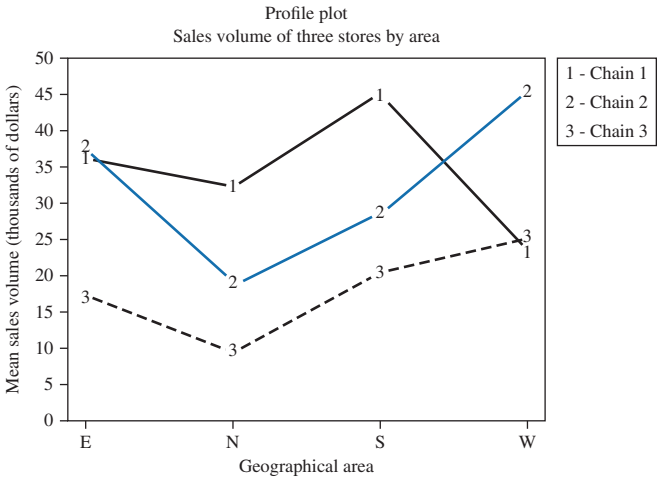
$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$

c. The AOV table is given here:

Source	df	SS	MS	*F*	*p*-value
Area	3	522.12	174.04	18.69	.0001
Chain	2	1,281.58	640.79	68.80	.0001
Area*chain	6	953.75	158.96	17.07	.0001
Week	1	22.04	22.04	2.37	.1519
Error	11	102.46	9.31		
Total	23	2,881.96			

There is significant evidence (p-value $< .0001$) of an interaction between area and chain. The profile plot displays an estimate of the type of interaction involved in the two factors.

The chain having the greatest mean sales volume changes from area to area.



ANSWER 19.27

19.27 **a.** The model for this situation is:

$y_{ijk} = \mu + \gamma_k + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$

where y_{ijk} is the sales volume during week k at chain i in area j,

γ_k is the effect of week k,

τ_i is the effect of chain i,

β_j is the effect of area j,

$\tau\beta_{ij}$ is the interaction effect of chain i in area j, and

ε_{ijk} is the random effect of all other factors.

b. To test for an interaction between area and chain, we would fit a reduced model with the interaction removed. Compute the difference in SSE between the reduced and complete models.

c. The complete model is given in part (a). The reduced model is

$y_{ijk} = \mu + \gamma_k + \tau_i + \beta_j + \varepsilon_{ijk}$

where the interaction is removed from the model.

If the interaction term is significant, then the test for main effects, in most situations, is not meaningful. If the interaction is found to be nonsignificant, then a test for main effect due to area can be conducted by fitting a reduced model with both the interaction term and the area main effect term deleted from the model. The complete model is now the model with both main effects but the interaction term removed. The reduced model is the model with both the interaction and the main effect due to area removed, but the main effect due to chain retained in the model. A similar procedure could be conducted to test for a main effect due to chain.

19.29 **a.** We can use a mixed-model approach to test the relevant hypotheses.

b. The interaction between training and inspector and the main effects due to training and inspector are the factors to be tested. We obtain the following test statistics:

Training*inspector: $F = \dfrac{MS_{T*I}}{MSE} = \dfrac{1.5/1}{106.33/16} = .23 \Rightarrow p\text{-value} = .6380 \Rightarrow$

There is not significant evidence of an interaction effect between inspectors and training.

Training: To determine the test statistic for testing the main effect due to training, we need to examine the expected MS column. We note that under the null hypothesis of no main effect due to training, $\theta_T = 0$. This implies that under the null hypothesis of no main effect due to training

$EMS_T = EMS_{L(T)} + EMS_{T*I} - EMSE$

Thus, the denominator of our test statistic is

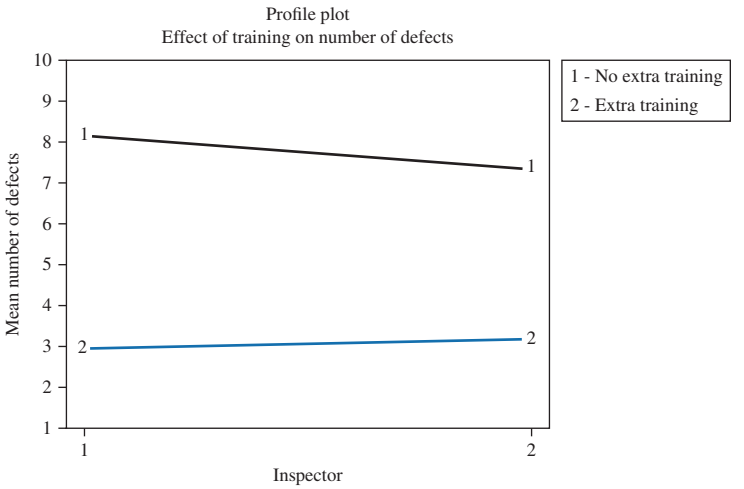
$M = MS_{L(T)} + MS_{T*I} - MSE = 14.17 + 1.5 - 6.65 = 9.02$. Using the Satterthwaite approximation, we obtain df $\geq$ 1.47. Therefore,

$$F = \frac{MS_T}{M} = \frac{130.67}{9.02} = 14.49 \text{ with } p\text{-value} = .0987$$

There is not significant evidence of an effect due to training. That is, the additional training does not appear to have reduced the mean number of defects.

Similarly, we determine there is not a significant effect due to inspectors (p-value = .6257).

c. A profile plot of the mean number of defects for the levels of training is give here:



REFERENCES

Chapter 1

Carroll, R., R. Chen, E. George, T. Li, H. Newton, H. Schmiediche, and N. Wang. (1997). "Ozone exposure and population density in Harris County, Texas." *Journal of the American Statistical Association* 92, 392–415.

Fontenot, B., L. Hunt, Z. Hildenbrand, D. Carlton, H. Oka, J. Walton, D. Hopkins, A. Osorio, B. Bjorndal, Q. Hu, and K. Schug. (2013). "An evaluation of water quality in private drinking water wells near natural gas extraction sites in the barnett shale formation." *Environmental Science Technology* 47(17), 10032–10040.

Hansen, L. P. (2006). "Migration and survival of farmed Atlantic salmon (Salmo salar L.) released from two Norwegian fish farms." *ICES Journal of Marine Science* 63, 1211–1217.

Hoynes, H. and Schanzenbach, D. (2012). " Work incentives and the food stamp program." *Journal of Public Economics* 96(1–2), 151–162.

"'Informaties' helps doctors unlock medical mysteries in mounds of data." *Houston Chronicle*, August 11, 2013.

National Research Council Committee. (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: National Academies Press.

Raftery, A., and J. Zeh. (1998). "Estimating bowhead whale population size and rate of increase from the 1993 census." *Journal of the American Statistical Association* 93, 451–462.

Rowson, S., and S. Duma. (2011). "Development of the STAR evaluation system for football helmets: integrating player head impact exposure and risk of concussion." *Annals of Biomedical Engineering* 39(8), 2130–2140.

Spiegelman, C., W. A. Tobin, W. D. James, S. J. Sheather, S. Wexler, and D. M. Roundhill. (2007). "Chemical and forensic analysis of JFK assassination bullet lots: Is a second shooter possible?" *Annals of Applied Statistics* 1, 287–301.

Utts, J. (2003). "What educated citizens should know about statistics and probability." *The American Statistician* 57, 74–79.

Chapter 2

Cryer, J., and R. Miller. (1991). *Statistics for Business: Data Analysis and Modelling.* Boston: PWS-Kent.

Freeman, S. F. (2004). *The Unexplained Exit Poll Discrepancy* (Research Report). Philadelphia University of Pennsylvania.

Lohr, S. (1999). *Sampling: Design and Analysis.* Pacific Grove, CA: Duxbury Press.

Scheaffer, R. L., W. Mendenhall, and L. Ott. (2006). *Elementary Survey Sampling.* 5th ed. Boston: Duxbury Press.

U.S. Bureau of Labor Statistics. (1997). *Handbook of Methods,* Vols. I and II. Washington, DC: U.S. Department of Labor.

Chapter 3

"From the Capital to the Classroom: Year 3 of the No Child Left Behind Act" Center on Education Policy, January 2004

National Commission on Excellence in Education. (1983). *A Nation at Risk: The Imperative for Education Reform.* Washington, DC: U.S. Government Printing Office.

Tekwe, C., R. Carter, C. Ma, J. Algina, M. Lucas, J. Roth, M. Ariet, T. Fisher, and M. Resnick. (2004). "An empirical comparison of statistical models for value-added assessment of school performance." *Journal of Educational and Behavioral Statistics* 29, 11–36.

Thall, P., and S. Vail. (1990). "Some covariance models for longitudinal count data with overdispersion." *Biometrics* 46, 657–671.

Tukey, J. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

U.S. Census Bureau. (2002). *Statistical Abstract of the United States.* 122nd ed. Washington, DC: U.S. Government Printing Office 2001.

U. S. Census Bureau. (2012). *Statistical Abstract of the United States*. Table 503, page 332.

Chapter 4

Barnard, G. A. (1958). "Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances:" Reproduced with the permission of the Council of

the Royal Society from *The Philosophical Transaction* (1763), 53, 370–418. Reprinted in 1956. *Biometrika* 45, 293–315.

Berry, D., and L. Chastain. (2004). "Inference about testosterone abuse among athletes." *Chance* 17, 5–8.

Devore, J. (2000). *Probability and Statistics for Engineering and the Sciences.* 5th ed. Boston: Duxbury Press.

Rao, P., J. Rhea, R. Novelline, A. Mostafavi, and C. McCabe. (1998). "Effect of computed tomography of the appendix on treatment of patients and use of hospital resources." *New England Journal of Medicine* 338, 141–146.

Repasky, R. (1991). "Temperature and the northern distributions of wintering birds." *Ecology* 72, 2274–2285.

Valway, S., M. Sanchez, T. Shinnick, I. Orme, T. Agerton, D. Hoy, S. Jones, H. Westmoreland, and I. Onorato. (1998). "An outbreak involving extensive transmission of a virulent strain of mycobacterium tuberculosis." *New England Journal of Medicine* 338, 633–639.

Chapter 5

Efron, B. (1979). "Bootstrap methods: another look at the jackknife." *Annals of Statistics* 7, 1–26.

Rosner, B., W. Willett, and D. Spiegelman. (1989). "Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error." *Statistics in Medicine* 8, 1051–1069.

Underwood, A., and J. Adler. (2004). "What you don't know about fat." *Newsweek*, August 23.

Williams, T., J. Rees, A. Ferguson, R. Herd, K. Kairu, and Y. Yobe. (1997). "Metals, petroleum hydrocarbons and organochlorines in inshore sediments and waters of Mombasa, Kenya." *Marine Pollution Bulletin* 34, 570–577.

Chapter 6

Conover, J. (1999). *Practical Nonparametric Statistics.* 3rd ed. New York: Wiley Press.

Ellis, J., R. Russell, F. Makrauer, E. Schaefer. (1987). "Increased risk for vitamin A toxicity in severe hypertriglyceridemia." *Annals of Internal Medicine* 105(6), 877–879.

Feder, H., and A. Blanchard. (1998). "The deep benthos of Prince William Sound, Alaska, 16 months after the Exxon Valdez oil spill." *Marine Pollution Bulletin* 36, 118–130.

Hughes, K., D. Weseloh, B. Braune. (1998). "The ratio of DDE to PCB concentrations in Great Lakes herring gull eggs and its use in interpreting contaminants data." *Journal of Great Lakes Research* 24(1), 12–31.

Miller, B. F., and C. B. Keane. (1957). *Encyclopedia of Medicine, Nursing, and Allied Health.* Philadelphia: W. B. Saunders.

Morton, D., A. Saah, S. Silberg, W. Owens, M. Roberts, M. Saah. (1982). "Lead absorption in children of employees in a lead industry." *American Journal of Epidemiology* 115, 549–555.

Murakimi, H., H. Oqawara, K. Morita, T. Saitoh, T. Matsushima, J. Tamura, M. Sawamura, M. Karasawa, S. Miyawaki, S. Schimano, S. Satoh, J. Tsuchiya. (1997). "Serum beta-2-microglobulin (SB2M) in patients with multiple myeloma treated with alpha interferon." *Journal of Medicine* 28, 311–318.

Newman, R. (1998). "Testing parallelism among the profiles after a certain time period." PhD diss., Texas A&M University.

Randles, R., and D. Wolfe. (1979). *Introduction to the Theory of Nonparametric Statistics.* New York: Wiley Press.

Santello, J., V. Dichtchekenian, J. Heimann. (1997). "Effect of long-term blood pressure control on salt sensitivity." *Journal of Medicine* 28, 147–156.

Welch, B. L. (1947). "The generalization of Student's problem when several population variances are involved." *Biometrika* 34, 28–35.

Chapter 7

Effron, B., and R. Tibshirani. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.

Manly, B. (1998). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 2nd ed. London: Chapman & Hall.

Pearson, E. S. and H. O. Hartley. (1976). "Biometrika tables for statisticians." 3rd ed. Cambridge: Cambridge University Press.

Power, C., S. McEwen, R. Johnson, M. Shourkri, K. Rahn, M. Griffiths, and S. De Grandis. (1998). "Repeatability of the Petrifilm HEC test and agreement with a hydrophobic grid membrane filtration method for the enumeration of Escherichia coli O157:H7 on beef carcasses." *Journal of Food Protection* 61, 402–408.

Chapter 8

Balli, S. J., J. F. Wedman, and D. H. Demo. (1997). "Family involvement with middle-grades homework: effects of differential prompting." *Journal of Experimental Education* 66, 31-48.

Box, G., and D. R. Cox. (1964). "An analysis of transformations." *Journal of Royal Statistical Society, Series B* 26, 211–252.

Draper, N., and H. Smith. (1998). *Applied Regression Analysis*. 3rd ed. New York: Wiley & Sons, Inc. vander Horst, C., P. Koster, C. de Borgie, P. Bossuyt, and M. van Gemert. (1998). "Effect of the timing of treatment of port-wine stains with the flash-lamp-pumped pulsed-dye laser." *New England Journal of Medicine* 338, 1028–1033.

Martx, H., P. Kvan, L. Abramson. (1996). "Empirical bayes estimation of the reliability of nuclear-power-plant emergency diescel generators." *Technometrics* 38, 11-24.

Wludyka, P., and P. Nelson. (1997). "An analysis-of-means-type test for variances from normal populations." *Technometrics* 39, 274–285.

Chapter 9

Cesare, S., R. Tannenbaum, A. Dalessio. (1990). "Interviewers' decisions related to applicant handicap type and rate empathy." *Human Performance* 3, 157–171.

Dunnett, C. (1955). "A multiple comparison procedure for comparing several treatments with a control." *Journal of the American Statistical Association* 50, 1096–1121.

Dunnett, C. (1964). "New tables for multiple comparisons with a control." *Biometrics* 20, 482–491.

Hollander, M., and D. Wolfe. (1999). *Nonparametric Statistical Methods*. 2nd ed. New York: Wiley-VCH.

Kramer, C. Y. (1956). "Extension of multiple range tests to group means with unequal numbers of replications." *Biometrics* 12, 307–310.

Krum, H., R. Viskoper, Y. Lacourcierre, M. Budde, and V. Charlton (1998). "The effect of an endothelin-receptor antagonist, Bosentan, on blood pressure in patients with essential hypertension." *The New England Journal of Medicine* 338, 784–791.

Scheffe, H. (1953). "A method for judging all contrasts in the analysis of variance." *Biometrika,* 40, 87–104.

Tukey, J. (1953). "The problem of multiple comparison." Manuscript.

Chapter 10

Agresti, A. (2002). *Categorical Data Analysis*. 2nd ed. New York: Wiley.

Agresti, A., and B. Coull. (1998). "Approximate is better than 'exact' for interval estimation of binomial proportions." *The American Statistician* 52, 119–126.

Bickel, P., E. Hammel, and J. O'Connell. (1975). "Sex bias in graduate admissions: Data from Berkeley." *Science* 187, 398–404.

Brown, L., T. Cai, and A. DasGupta. (2001). "Interval estimation for a binomial proportion." *Statistical Science* 16, 101–138.

Chen, S-H, and W-Y Chen. (1995). "Generalized minimal distortion segmentation for ANN-based speech recognition." *IEEE Trans. on Speech and Audio Processinq* 3(2), 141–145.

Cochran, W. (1954). "Some methods for strengthening the common χ^2 test." *Biometrics* 10, 417–451.

Conover, J. (1999). *Practical Nonparametric Statistics.* 3rd ed. New York: Wiley Press.

Fernandez-Cornejo, J. (2008). "Environmental and economic consequences of technology adoption: IPM in viticulture." *Journal of American Statistical Association* 84, 851–861.

Golombok, S., and F. Tasker. (1996). "Do parents influence the sexual orientation of their children?" *Developmental Psychology* 32(1), 3–11.

Hand, D., F. Daly, A. Lunn, K. McConway, and E. Ostrowski. (1993). *A Handbook of Small Data Sets*. London: Chapman & Hall.

Koehler, K. (1986). "Goodness-of-fit tests for log-linear models in sparse contingency tables." *Journal of the American Statistical Association* 81, 483–493.

Larntz, K. (1978). "Small-sample comparison of exact levels for chi-squared goodness-of-fit statistics." *Journal of the American Statistical Association* 73, 253–263.

Mantel, N., and W. Haenszel. (1959). "Statistical aspects of the analysis of data from retrospective studies of disease." *Journal of National Cancer Institute* 22, 719–748.

Margolin, B., B. S. Kim, and K. Risko (1989). "The Ames salmonell/microsome mutagenicity assay: issues of inference and validation." *Agricultural Economics* 18, 145–155.

McNemar, Q. (1947). "Note on the sampling error of the differences between correlated proportions or percentage." *Psucliometrika* 12, 153–157.

Meehan, W., R. Mannix, M. O'Brien, and M. Collins. (2013). "The prevalence of undiagnosed concussions in athletes." *Clinical Journal of Sports Medicine* 23(5), 339–342.

Meyer, M. and T. Finney (2005). "Who wants airbags." *Chance* 18, 3–16.

National Coalition for Women and Girls in Education (2002). "Title IX at 30, Report Card on Equity."

Raley, R., M. Frisco, E. Wildsmith (2005). "Maternal cohabitation and educational success." *Sociology of Education* 78, 144–164.

Tarlow, B., S. Wisniewski, S. Belle, M. Rubert, M. Ory, and D. Gallagher-Thompson (2004). "Positive aspects of caregiving." *Research on Aging* 26, 429–453.

Wilson, E. (1927). "Probable inference, the law of succession, and statistical inference." *Journal of American Statistical Association* 22, 209–212.

Chapter 11

Carter, R. (1981). "Restricted maximum likelihood estimation of bias and reliability in the comparison of several measuring methods." *Biometrics* 37, 733–741.

Chapter 12

Abraham, B., and J. Ledolter. (2006). *Introduction to Regression Modeling*. Belmont, CA: Thomson Brooks/Cole.

Der, G., and B. Everitt. (2002). *A Handbook of Statistical Analyses Using SAS.* 2nd ed. New York: Chapman & Hall/CRC.

Sheather, S. (2009). *A Modern Approach to Regression with R,* New York: Springer-Verlag.

Smith, A. F. (1967). "Diagnostic value of serum-creatinine-kinase in a coronary care unit." *Lancet* 290, 178–182.

Chapter 13

Abraham, B., and J. Ledolter. (2006). *Introduction to Regression Modeling*. Belmont, CA: Thomson Brooks/Cole.

Belsley, D., E. Kuh, and R. Welsch. (1980). *Regression Diagnostics.* New York: Wiley.

Box, G., and D. R. Cox. (1964). "An analysis of transformations." *Journal of Royal Statistical Society, Series B* 26, 211–252.

Brown, S., M. Healy, and M. Kearns. (1981). "Report on the interlaboratory trial of the reference method for the determination of total calcium in serum." *Journal of Clinical Chemistry and Clinical Biochemistry* 19, 395–426.

Brownlee, K. (1965). *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley.

Carroll, R., and D. Ruppert. (1988). *Transformation and Weighting in Regression*. London: Chapman & Hall.

Chatterjee, S., and A. Hadi. (2012). *Regression Analysis by Example*. New York: Wiley.

Cook, R. D., and S. Weisberg. (1982). *Residuals and Influence in Regression*. London: Chapman & Hall.

Cox, D., and E. Snell. (1981). *Applied Statistics: Principles and Examples.* London: Chapman & Hall.

Draper, N., and H. Smith. (1997). *Applied Regression Analysis*. 3rd ed. New York: Wiley.

Durbin, J., and G. Watson. (1951). "Testing for serial correlation in least squares, II." *Biometrika* 38, 159–178.

Hand, D., F. Daly, A. Lunn, K. McConway, and E. Ostrowski. (1993). *A Handbook of Small Data Sets*. London: Chapman & Hall.

Lange, T., H. Royals, and L. Connor. (1993). "Influence of water chemistry on mercury concentration in largemouth bass from Florida lakes." *Transactions of the American Fisheries Society* 122, 74–84.

Mallows, C. (1973). "Some comments on C_p." *Technometrics* 15, 661–675.

Neter, J., M. Kutner, C. Nachtsheim, and W. Wasserman. (1996). *Applied Linear Statistical Models.* 4th ed. Boston: WCB McGraw-Hill.

Peck, R., C. Olson, and J. Devore. (2005). *Introduction to Statistics and Data Analysis*. Belmont, CA: Thomson Brooks/Cole.

Sheather, S. (2009). *A Modern Approach to Regression with R*. New York: Springer.

Sokal, R., and F. J. Rohlf. (1981). *Biometry*. San Francisco: W. H. Freeman.

Tukey, J. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Chapter 14

Chin, K. B., J. T. Keeton, M. T. Longnecker, and J. W. Lamkey. (1999). "Utilization of soy protein isolate and konjac blends in a low-fat bologna (model system)." *Meat Science* 53, 45–57.

Fisher, R. A. (1951). *The Design of Experiments.* 6th ed. Edinburgh: Oliver and Boyd.

Kuehl, R. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*. 2nd ed. Boston: Duxbury Press.

Millikin, G., and D. Johnson. (1992). *Analysis of Messy Data*. Vol. 1 of *Designed Experiments.* New York: Chapman & Hall.

Chapter 15

Damaser, E., R. Shor, and M. Orne. (1963). "Physiological effect during hypnotically requested emotions." *Psychosomatic Medicine* 25, 334–343.

Hand, D., F. Daly, A. Lunn, K. McConway, and E. Ostrowski. (1993). *A Handbook of Small Data Sets*. London: Chapman & Hall.

Hollander, M., and D. Wolfe. (1999). *Nonparametric Statistical Methods*. 2nd ed. New York: Wiley-VCH.

Mason, R., R. Gunst, and J. Hess (2003). *Statistical Design and Analysis of Experiments*. 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc.

Chapter 16

Pyke, C., R. Condit, S. Aguilar, and S. Lao. (2001). "Floristic composition across a climatic gradient in a neotropicallowland forest." *Journal of Vegetation Science* 12, 553–566.

Chapter 17

Kuehl, R. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*. 2nd ed. Boston: Duxbury Press.

Oehlert, G. (2000). *A First Course in Design and Analysis of Experiments*. New York: W. H. Freeman and Company.

Searle, S., G. Casella, and C. McCulloch. (1992). *Variance Components.* New York: Wiley.

Chapter 18

Chinchilli, V., B. Schwab, P. K. Sen. (1989). "Inferences based on ranks ofr the multiple-design multivariate linear model." *Journal of American Statistical Association* 84, 517-524.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: Wiley.

Crowder, M., and D. Hand. (1990). *Analysis of Repeated Measures*. London: Chapman & Hall.

Diggle, P., K. Liang, and S. Zeger. (1996). *Analysis of Longitudinal Data.* New York: Oxford University Press.

Gennings, C., V. Chinchilli, W. Carter. (1989). "Response surface analysis with correlated data: a nonlinear approach." *Journal of American Statistical Association* 84, 305-309.

Greenhouse, S., and S. Geisser. (1959). "On methods in the analysis of profile data." *Psychometrika* 24, 95–112.

Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge: Cambridge University Press.

Huynh, H., and L. Feldt. (1970). "Conditions under which mean square ratios in repeated measurement designs have fixed F-distributions." *Journal of the American Statistical Association* 65, 1582–1589.

Jones, B., and M. Kenard. (2015). *Design and Analysis of Cross-Over Trials*. London: Chapman & Hall.

Kuehl, R. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis.* 2nd ed. Boston: Duxbury Press.

Newman, R. (1998). "Testing parallelism among the profiles after a certain time period." PhD diss. Texas A&M University.

Oehlert, G. (2000). *A First Course in Design and Analysis of Experiments.* New York: W. H. Freeman and Company.

Ripley, B. (1988). *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Snedecor, G., and W. Cochran. (1980). *Statistical Methods.* 7th ed. Ames, IA: Iowa State University Press.

Vonesh, E., and V. Chinchilli. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker, Inc.

Chapter 19

Cochran, W., and G. Cox. (1957). *Experimental Design*. 2nd ed. New York: Wiley.

Kuehl, R. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*. 2nd ed. Boston: Duxbury Press.

Lenter, M., and T. Bishop. (1993). *Experimental Design and Analysis*. 2nd ed. Blacksburg, VA: Valley Book Company.

Little, R., and D. Rubin. (2002). *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley.

INDEX

A

B

C

D

E

H

I

K

L

M

Q

R

S

T

U

V

W

Y

Z