D. H. Stamatis

SIX SIGMA AND BEYOND

Statistics and Probability

St. Lucie Press

SIX SIGMA AND BEYOND

Statistics and Probability

SIX SIGMA AND BEYOND

A series by D.H. Stamatis

Volume I
Foundations of Excellent Performance

Volume II
Problem Solving and Basic Mathematics

Volume III
Statistics and Probability

Volume IV
Statistical Process Control

Volume V
Design of Experiments

Volume VI
Design for Six Sigma

Volume VII
The Implementation Process

D. H. Stamatis

SIX SIGMA AND BEYOND

Statistics and Probability

St. Lucie Press

A CRC Press Company
Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Stamatis, D. H., 1947-
Six sigma and beyond: statistics and probability, volume III
p. cm. — (Six Sigma and beyond series)
Includes bibliographical references and index.
ISBN 1-57444-3127
1. Quality control—Statistical methods. 2. Production management—Statistical methods. 3. Industrial management. I. Title. II. Series.

TS156.S73 2001
658.5'62—dc21 2001041635
CIP

International Standard Book Number 1-57444-3127
Library of Congress Card Number 2001041635
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

To Stephen

Preface

> The long-range contribution of statistics depends not so much upon getting a lot of highly trained statisticians into industry, as it does in creating a statistically minded generation of physicists, chemists, engineers, and others who will in any way have a hand in developing and directing the production processes of tomorrow.
>
> W.A. Shewhart and W.E. Deming

Much has been said about statistics and their use. Often, though, we statisticians overlook the discussion of the obvious as soon as we move away from the academic arena. We expect researchers and professionals in all walks of life to use the many tools offered by the statistical world, but we have failed to educate them appropriately both in concept and application. The focus of most statistics books seems to be formula utilization.

This volume will attempt to explain the tools of statistics and to provide guidance on how to use them appropriately and effectively. The structure of this work is going to follow (1) the conceptual domain of some useful statistical tools, (2) appropriate formulas for specific tools, and (3) the connection between statistics and probability.

This volume is not intended to be a textbook. It is intended to be a general manual for people who are interested in using statistical, probability, and reliability concepts to improve processes and profitability in their organizations.

The discussion begins with very elementary issues and progresses to some very advanced tools for decision-making. Specifically, the book begins by delineating the importance of collecting, analyzing, and interpreting data, from a practical perspective rather than an academic point of view. The assumption is that you (the reader) are about to begin a study of something, and you want to do it well. You want to design a good study, analyze the results properly, and prepare a cogent report that summarizes what you have found.

Because of these assumptions, this book does not dwell on formulas and significance tables or proofs for that matter. The assumption is that a statistical software package will be utilized, and that the reader will benefit more from learning to understand and interpret the results generated by that software than from memorizing formulas.

About the Author

D. H. Stamatis, Ph.D., ASQC-Fellow, CQE, CMfgE, is currently president of Contemporary Consultants, in Southgate, Michigan. He received his B.S. and B.A. degrees in marketing from Wayne State University, his Master's degree from Central Michigan University, and his Ph.D. degree in instructional technology and business/statistics from Wayne State University.

Dr. Stamatis is a certified quality engineer for the American Society of Quality Control, a certified manufacturing engineer for the Society of Manufacturing Engineers, and a graduate of BSIIs ISO 9000 lead assessor training program.

He is a specialist in management consulting, organizational development, and quality science and has taught these subjects at Central Michigan University, the University of Michigan, and Florida Institute of Technology.

With more than 30 years of experience in management, quality training, and consulting, Dr. Stamatis has served and consulted for numerous industries in the private and public sectors. His consulting extends across the United States, Southeast Asia, Japan, China, India, and Europe. Dr. Stamatis has written more than 60 articles and presented many speeches at national and international conferences on quality. He is a contributing author in several books and the sole author of 12 books. In addition, he has performed more than 100 automotive-related audits and 25 preassessment ISO 9000 audits, and has helped several companies attain certification. He is an active member of the Detroit Engineering Society, the American Society for Training and Development, the American Marketing Association, and the American Research Association, and a fellow of the American Society for Quality Control.

Acknowledgments

In a typical book, the author begins by thanking several individuals who have helped to complete it. In this mammoth work, so many people have helped that I am concerned that I may forget someone.

The writing of a book is a collective undertaking by many people. To write a book that conveys hundreds of thoughts, principles, and ways of doing things is truly a Herculean task for one individual. Since I am definitely not a Hercules or a Superman, I have depended on many people over the years to guide me and help me formulate my thoughts and opinions about many things, including this work. To thank everyone by name who has contributed to this work would be impossible, although I am indebted to all of them for their contributions. However, some organizations and individuals do stand beyond the rest, and without them, this series would not be possible.

Special thanks go to Dr. A. Stuart for granting me permission to use and adopt much of the discussion on discrete random variables, continuous RVs, uniform and beta distributions, functions of random variables (tolerances), exponential distribution and reliability, and hypothesis testing and OC curves in Part II of this volume. The work was adapted from the notes of *Statistics and Probability for Engineers* used as training material at Ford Motor Company.

Special thanks also go to Duxburry Press for granting me permission to use the material on Holt's Model for trend and Winters' Model for seasonality and econometric models. The work is based on *Managerial Statistics* by S.C. Albright, W.L. Winston, C.J. Zappe and P. Kolesar, published in 2001.

In addition, special thanks go to Prentice Hall for granting me permission to use the material on the summary of differences between MANOVA and discriminant analysis, what is conjoint analysis, uses of conjoint analysis, what is canonical correlation, and what is cluster analysis. The work is based on *Multivariate Data Analysis,* 5th ed., by J.F. Hair, R.E. Anderson, R.L. Tathan, and W.C. Black, published in 1998.

I would like to thank my colleagues Dr. R. Rosa, H. Jamal, Dr. A. Crocker, and Dr. D. Demis, as well as J. Stewart and R. Start, for their countless hours of discussions in formulating the content of these volumes in their final format.

In addition, I want to thank J. Malicki, C. Robinson, and S. Stamatis for their computer work in preparing some of the earlier drafts and final figures in the text.

I would like to thank as always my personal inspiration, bouncing board, navigator and editor, Carla, for her continually enthusiastic attitude during my most trying times. Especially for this work she has demonstrated extraordinary patience, encouragement, and understanding in putting up with me.

Special thanks go to the editors of the series for their suggestions and improvements of both the text and its presentation in the final format.

Finally, my greatest appreciation is reserved for my seminar participants and the students of Central Michigan University who, through their input, concerns, and discussions, have helped me to formulate these volumes. Without their active participation and comments, these volumes would never have been finished. I really appreciate their effort.

List of Figures

List of Tables

Table of Contents

PART I Essential Concepts of Statistics

PART III Appendices

Part I

Essential Concepts of Statistics

Introduction

This introduction will discuss the basic concepts of all statistics. The intent of the introduction is to sensitize the reader to the importance of taking statistics into consideration in the design and planning of experiments. Unless the experimenter plans a study appropriately, accounts for certain issues that are inherent in any study, and understands what is needed for a successful experiment, all will be for naught.

WHAT ARE DATA?

Everything we do is based on data. So, the question quite often is: should the word be datum or data? Grammatically speaking, the singular word is datum and the plural is data. However, because generally speaking we have more than one, the convention is that we use data. In common usage, data are any materials that serve as a basis for drawing conclusions. (Notice that the word we use is "materials." That is because materials may be quantifiable or numerical and measurable or on the other hand may be attribute or qualitative. In either case they can be used for drawing conclusions.) Drawing conclusions from data is an activity in which everyone engages — bankers, scholars, politicians, doctors, and corporate presidents. In theory, we base our foreign policy, methods of treating diseases, corporate marketing strategies, and process efficiency and quality on "data."

Data come from many sources. We can conduct our own surveys or experiments, look at information from surveys other people have conducted, or examine data from all sorts of existing records — such as stock transactions, election tallies, or inspection records. But acquiring data is not enough. We must determine what conclusions are justified based on the data. That is known as "data analysis." People and organizations deal with data in many different ways. Some people accumulate data but do not bother to evaluate it objectively. They think that they know the answers before they start. Others want to examine the data but do not know where to begin. Sometimes people carefully analyze data, but the data are inappropriate for the conclusions that they want to draw. Unless the data are correctly analyzed, the "conclusions" based on them may be in error. A superior treatment for a disease may be dismissed as ineffectual; you may purchase stocks that do not perform well and lose your life's savings; you may target your marketing campaign to the wrong audience, costing your company millions of dollars; or you may adjust the wrong item in a process, and as a consequence, you may affect the response of the customer

in a very unexpected way. The consequences of bad data analysis can be severe and far-reaching. That is why you need to know how to analyze data well.

You can analyze data in many different ways. Sometimes all you need to do is describe the data. For example, how many people say they are going to buy a new product you are introducing? What proportion of them are men and what proportion are women? What is their average income? What product characteristic is the customer delighted with? In other situations, you want to draw more far-reaching conclusions based on the data you have at hand. You want to know whether your candidate stands a chance of winning an election, whether a new drug is better than the one usually used, or how to improve the design of a product so that the customer will be really excited about it. You do not have all of the information you would like to have. You have data from some people or samples, but you would like to draw conclusions about a much larger audience or population.

At this juncture your answer may be, "I do not have to worry about all this because the computer will do it for me." That is not an absolute truth. Computers simplify many tasks, including data analysis. By using a computer to analyze your data, you greatly reduce both the possibility of error and the time required. Learning about computers and preparing data for analysis by computer do require time, but in the long run they substantially decrease the time and effort required. Using a computer also makes learning about data analysis much easier. You do not have to spend time learning formulas. The computer can do the calculating for you. Instead, your effort can go into the more interesting components of data analysis — generating ideas, choosing analyses, and interpreting their results.

Because calculations are the computer's job, not yours, this volume does not emphasize formulas. It emphasizes understanding the concepts underlying data analysis. The computer can be used to calculate results. You need to learn how to interpret them.

DESCRIBING DATA

Once you have prepared a data file, you are ready to start analyzing the data. The first step in data analysis is describing the data. You look at the information you have gathered and summarize it in various ways. You count the number of people giving each of the possible responses. You describe the values by calculating averages and seeing how much the responses vary. You look at several characteristics together. How many men and how many women are satisfied with your new product? What are their average ages? You also identify values that appear to be unusual, such as ages in the one hundreds or incomes in the millions, and you check the original records to make sure that these values were picked up correctly. You do not want to waste time analyzing incorrect data.

TESTING HYPOTHESES

Sometimes you have information available for everyone or everything that you are interested in drawing conclusions about, and all you need to do is summarize your

data. But usually that is not the case. Instead, you usually want to draw conclusions about much larger groups of people or objects than those included in your study. You want to know what proportion of all purchasers of your product are satisfied with it, based on the opinions of the relatively small number of purchasers included in your survey. You want to know whether buyers of your product differ from nonbuyers. Are they younger, richer, better educated? You want to be able to draw conclusions about all buyers and nonbuyers based on the people you have included in your study.

To do this (and understand it), you have to learn something about statistical inference. Later chapters in this volume will show you how to test hypotheses and draw conclusions about populations based on samples. You will learn how to test whether you have sufficient evidence to believe that the differences or relationships you find in your sample are true for the whole population.

DESCRIBING RELATIONSHIPS

You often want to determine what the relationship is between two variables. For example, what is the relationship between dollars spent on advertising and sales? How can you predict how many additional sales to expect if you increase your advertising budget by 25%? What is the relationship between the dosage of a drug and the reduction in blood pressure? How can you predict the effect on blood pressure if you cut the dose in half? You can study and model the relationship between pairs of variables in many different ways. You can compute indexes that estimate the strength of the relationship. You can build a model that allows you to predict values of one variable based on the values of another. That is what the last part of the book is about.

You must state your ideas clearly if you plan to evaluate them. This advice applies to any kind of work but especially to research design and statistical analysis. Before you begin working on design and analysis, you need to have a clearly defined topic to investigate.

ASKING A QUESTION

You may have a general suspicion that smoking less makes people feel better. You may think that component A is better than component B. Or you may have an idea for a study method that will make people learn more. Before you begin a study about such intuitions, you should replace vague concepts such as “feeling better” or “smoking less” or “learning more” with definitions that describe measurements that you can make and compare. You might define “better” with a specific performance improvement or a reduction in failure. You might replace “feeling better” with an objective definition such as “the subject experiences no pain for a week.” Or you might record the actual dosage of medication required to control pain. If you are interested in smoking, you need a lot of information to describe it. What does each of the subjects smoke — a pipe, cigars, or cigarettes? How much tobacco do the subjects use in a day? How long have they been smoking? Has the number of cigarettes (or cigars or pipes) that they smoke changed?

On the other hand, you must balance your scientific curiosity with the practical problems of obtaining information. If you must rely on people's memory, you cannot ask questions like "What did you have for dinner ten years ago?" You must ask questions that people will be able to answer accurately. If you are trying to show a relationship between diet and disease, for example, you cannot rely on people's memory of what they ate at individual meals. Instead, you have to be satisfied with overall patterns that people can recall. Some information is simply not available to you, however much you would like to have it. It is better to recognize this fact before you begin a study than when you get your questionnaires back and find that people were not able to answer your favorite question. If you think about your topic in advance, you can substitute a better question — one that will give you information you can use, even if it is not the information you wish you could have.

WHAT INFORMATION DO YOU NEED?

A critical step in the design of any study is the decision about what information you are going to record. Of course, you cannot record every possible piece of information about your subjects and their environment. Therefore, you should think hard about what information you will try to get. If you accidentally forget to find out about an important characteristic of your subjects, you may be unable to make sense of the patterns you find in your data. When in doubt, it is usually better to record more information than less. It is easy to leave unnecessary variables out of your data analysis, but it is often difficult (and expensive) to go back and gather additional information. For example, if you are studying what types of people are likely to buy a high-priced new product, you may not be able to adequately compare buyers with nonbuyers if you forget to include information about income.

DEFINING A POPULATION

When you conduct a study, you want your conclusions to be far-reaching. If you are a psychology student, you may want your results to apply to all laboratory rats, not just the ones in your lab. Similarly, if you are doing a market research survey on whether people in Los Angeles would buy disposable umbrellas, you may want to draw conclusions about everybody in the city. If you are an engineer and you are involved in the development of a particular product, you want to know what kind of a base or population the product is for. The people or objects about whom you want to draw conclusions are called a *population.*

One of the early steps in any study is nailing down exactly what you want your population to be. The more definite you are in defining populations, the better your understanding of samples and the results of your study will be.

Defining a population may seem straightforward, but often it is not. Suppose that you are a company personnel manager, and you want to study why people miss work. You probably want to draw conclusions only about employees in your particular company. Your population is well defined. However, if you are a graduate student writing a dissertation about the same topic, you face a much more complicated

problem. Do you want to draw conclusions about professionals, laborers, or clerical staff? About men or women? Which part of the world is of interest — a city, a country, or the world as a whole? No doubt, you (and your advisor) would be delighted if you could come up with an explanation for absenteeism that would apply to all sorts of workers in all sorts of places. You are not likely to come up with that kind of explanation, though. Even if you do, you are not likely to come up with the evidence to support it.

All kinds of people miss work because they are sick, but unlike others, the president of Major Corporation probably does not need to stay home waiting for a phone to be installed. The afternoons he takes off to play golf with his buddies are probably not recorded by the personnel office as absenteeism, either. People miss work for lots of reasons, and the reasons are quite different for different kinds of employees. Be realistic and study only a part of the labor force. Absenteeism among laborers in auto factories in Detroit, for example, is a problem with a well-defined population about which you would have a fighting chance to draw some interesting conclusions.

DESIGNING A STUDY

Even when the population of interest seems to be well defined, you may not actually be able to study it. If you are evaluating a new method for weight loss, you would ideally like to draw conclusions about how well it works for all overweight people. You cannot really study all overweight people, though, or even a group that is typical of all overweight people. People who do not want to lose weight or who have been disheartened by past efforts to reduce may not agree to try yet another method. You will probably be able to try out your new method only on people who want to lose weight and who have not given up trying. These people, not all overweight people, form your population.

Remember that a population defined realistically in this way may be different from the ideal population. For example, the population in your weight loss study may be lighter, younger, or healthier than the ideal population of all overweight people. Therefore, your conclusions from studying people who want to lose weight do not necessarily apply to people who are not motivated. For example, the treatment may have some unpleasant consequences, such as making people want to chew on the nearest thing available, such as gum, a pencil, or the corner of a desk. People who really want to lose weight may be willing to put up with such minor inconveniences in order to reach their goal. People who do not care much about their weight probably will not be. Thus, the new treatment may work quite differently for those who are motivated versus those who are not.

SAMPLING

Although you may want to draw conclusions about all rats or all residents of Los Angeles, you certainly do not want to have to train all of the world's rats or personally visit every Los Angeles home. What you want to do is to study some rats or some

people, draw conclusions based on what you have observed for them, and have the conclusions apply to the population in which you are really interested.

The rats or people (or other creatures or objects) that you actually observe in your study are called the *sample.* You can select a sample from a particular population in countless ways. How you do it is very important because if you do not do it correctly, you will not be able to draw conclusions about your population. That is a pretty serious shortcoming. For the most part, interesting studies are those that allow you to draw conclusions about a much larger group of subjects than that actually included in the sample.

RANDOM SAMPLES

What is a good sample? A sample is supposed to let you draw conclusions about the population from which it is taken. Therefore, a good sample is one that is similar to the population you are studying. But you should not go out and just look for animals, vegetables, or minerals that you think are "typical" of your population. With that kind of a sample (a judgment sample), the reliability of the conclusions you draw depends on how good your judgment was in selecting the sample — and you cannot assess the selection scientifically. If you want to back up your research judgments with statistics (one of the reasons, I hope, why you are reading this book), you need a *random sample.* Statisticians have studied the behavior of random samples thoroughly. As you will learn in later chapters, the very fact that a sample is random means that you can determine what conclusions about the population you can reasonably draw from the sample.

So what is a random sample, if it is so important? It is a sample that gives every member of the population (animal, vegetable, mineral, or whatever) a fair chance of selection. Everyone or everything in the population has the same chance. No particular type of creature or thing is systematically *excluded* from the study, and no particular type is more likely than any other to be *included.* Also, each unit is selected independently: including one particular unit does not affect the chance of including another.

If you are interested in the opinions of all the adults in Los Angeles, do not rely on a door-to-door poll in mid-afternoon or ask questions of people as they leave church services on a rainy Sunday. Such samples exclude many of the types of people you want to draw conclusions about. People who have jobs are usually not home on weekday afternoons, so their opinions would not be included in your results. Similarly, people standing in the rain may express different opinions (especially about umbrellas, for example) than they would if they were warm and dry. Polling in the rain would lead you to a bad guess about the proportion of the city's residents interested in your new product (disposable umbrellas). To make things worse, you cannot tell what the effects of excluding dry people will be. You cannot tell whether your observed results are biased one way or another, and you cannot tell by how much. You might even be on target, but you do not know that, either.

From any particular random sample, of course, the results are not exactly the same as the results you would get if you included the entire population. Later chapters will show you how statistical methods take into account the fact that different

samples lead to somewhat different results. You will then understand how much you can say about a population from the results you observe in a sample.

VOLUNTEERS

To make it easier to have people participate in your study, you may be tempted to rely on volunteers. But you should not rely on any special types of people, and volunteers are one of those special types. Many studies have shown that people who volunteer differ in important ways from those who do not.

By the same token, if you are interested in testing a particular product, you should not base your decisions only on bad samples just because they have failed. You do not know enough yet about the causes of the failure or the conditions under which it occurred. Conversely, you do not test only good samples because they have no failures. In both cases the results will be erroneous.

USING SURVEYS

Generally, there are two major categories of studies: (1) surveys and (2) experiments. Other categories of studies also exist, but these two are predominant. The two types of studies differ in important ways.

In a survey, one records information. You ask people questions and record their answers, or you take some kind of a measurement. The important thing is that the experimenter does not actually do anything to the subjects or objects of the study. In fact, the experimenter tries very hard not to exert any influence whatsoever.

To conduct a good survey, the experimenter must phrase the questions so they do not suggest "correct" answers. In the case of surveying products, the experimenter must be conscious of their location, category, and so on, so that a general profile may be reconstructed with the results obtained and not by limited selection or discrimination of the product.

The great advantage to conducting your own survey is that you can tailor it for your own research project. You can ask the questions you want to ask in the way you want to ask them. You can choose the exact population that you want to study and select just the kind of sample you need. You can control the training of interviewers, and you can deal with all of the problems that come up during the actual survey. In short, you can do everything possible to make sure the survey will help you answer your specific questions of interest.

Doing all of these things takes a great deal of time and often a great deal of money. If you are going to invest a lot of time and money in a study, you owe it to yourself to get expert advice. Show your plans to someone who has actually carried out similar surveys, and ask for advice — *before* you take any big steps such as printing the questionnaires. If in doubt, consult a statistician or a book on data analysis.

ANALYZING AN EXISTING SURVEY

Without a doubt, the best way to get survey data is to design and carry out a survey focused on precisely the research questions you want to study. Realistically, though,

you often have to settle for "re-using" a survey that somebody else has carried out. Using data from a survey that was not designed for your study is often called *secondary analysis* to distinguish it from the *primary analysis* that was the purpose of the original survey.

Secondary analysis lets you do research that you could not otherwise do all on your own. But you must keep in mind that the data were not collected specifically for your purposes. The survey questions may not have measured exactly what you wanted them to, but you are stuck with them nonetheless. Remember to interpret them as they were asked, not as you wish they had been asked.

When you plan to use existing data, you do not have to worry about the thousands of details that go into conducting a survey. Instead, you have to make sure that the survey was carried out properly in the first place. Was it conducted by a reputable organization? Were the questions well phrased? Was the sample well chosen? Were the forms carefully processed? Most important, have you formulated research questions that you can reasonably hope to answer with the existing data?

DESIGNING EXPERIMENTS

Unlike a survey, an experiment involves actually doing something to the subjects or objects rather than just soliciting answers to questions or making measurements. For example, instead of asking people whether they think that vitamin C is effective for preventing colds, you might give them vitamin C and observe how many colds they develop. Or you may want to try product A and product B and then compare the results to see which one is better. Sometimes you study the subjects before and after your experimental treatment. Sometimes, instead, you take several groups of subjects, do something different to each of the groups, and then compare the results.

Experimentation on people poses ethical questions that deserve careful thought. Many responsible institutions have committees that regulate experiments involving human subjects. If an experiment exposes a subject to risks, such as possible side effects from a new drug, you must certainly inform the subjects in advance. Usually you must have them sign forms to give their consent. Needless to say, that is not a concern when you test products — even though the test may be a destructive one.

In experiments as well as surveys, the subjects must come from the population that you are interested in. (As you have probably gathered by now, proper sampling is much easier with animals, processes, or products in a laboratory setting than with people in a survey or products in a real world application.) When you design an experiment, you need to fret about some other things as well. For example, to compare different treatments or techniques, you must make sure that the groups receiving them are as similar as possible. Again, randomness is the key. The best way to make groups similar is to assign subjects or objects to the groups randomly. This procedure does not guarantee that the groups will be exactly the same, but it does increase the likelihood.

RANDOM ASSIGNMENT

Random does not mean "any old way." You cannot assign subjects or objects to groups according to whatever strikes your fancy or let others make the assignment

TABLE I.1
A Limited Table of Random Numbers

8588	5171	0775	7818	8683	3168 …
7185	8645	1537	3754	0201	2450 …
1053	9728	3028	8725	4855	0218 …
7517	0826	7257	5527	2668	8157 …
3551	3316	3584	9439	0011	7365 …
⋮					
7405	7764	6131	6204	8835	0345 …

decisions for you. Randomness requires a very specific, systematic approach to minimize the chance of distortion of groups due to the inclusion of disproportionate numbers of particular types of individuals or products.

If you allowed teachers to select which of their students receive personal computers, for example, they might select well-behaved students to reward them for past efforts. These students may be more intelligent or more diligent than the students who do not get to use the special equipment. Any evaluation of the effect of personal computers would be tainted by the differences between the selected students and the population as a whole.

Or consider this example: An engineer is trying to study customers' perceptions of the effect of adjustable brakes in vehicles. The results would be very different if the sample was based only on individuals with a height of more than 5 feet 11 inches, rather than a random sample of drivers of different heights.

A good way to assign people, animals, or objects to groups is to use a table of random numbers. You cannot just make up a table of numbers that you think are random. You are likely to have certain number biases. Unlike experimenters, random number tables do not have birthdays, license plates, children, or any other reasons to prefer one number over another. In a properly constructed table of random numbers, every number from zero to nine has the same chance of appearing in any position in the table.

Table I.1 shows a small random number table. The table has the numbers grouped into fours, but the grouping is just for convenience. It has no other significance. To randomly assign subjects to groups, you start at an arbitrary place in the table and assign the digit at that place to the first subject or object. Each new subject gets a digit from successive places in the table. If you start at the fourth digit of the first vertical group in the fourth position in Table I.1, for example, and then proceed to the right, the first subject gets the number 7, the next subject the number 0, and the next subject the number 8. (These three digits have been printed in bold type.) Since everything is random, it really does not matter whether you read the table across or down. However, once you have selected a starting point, stay in sequence. Using the table in this systematic way prevents you from choosing "favorite" numbers as starting points or as the next numbers in the sequence. You can never be too careful when you are trying to be random.

You use the numbers you assigned to the subjects to assign them to experimental groups. For example, if you have two groups, you can assign subjects with even

numbers to one group and subjects with odd numbers to another group. This procedure should result in about the same number of subjects in the two groups. But if you want the groups to be exactly equal in size, you can assign two- or three-digit random numbers to each of the subjects. Then arrange the numbers in order, from smallest to largest. Subjects with numbers in the lower half go to one group, and subjects with numbers in the upper half go to the other. You can use all sorts of systems with a random number table to assign subjects to groups, even in very complicated experimental designs. It is customary nowadays to use a computer generator program to generate random numbers.

Does randomness really matter? Yes, it does. Unless you use a procedure that assigns your subjects randomly, the results of your study may be difficult or impossible to interpret. Many assignment schemes that appear random to the inexperienced investigator turn out to have hidden flaws. For example, on one occasion, researchers at a hospital compared two treatments for a particular disease. Patients who were admitted on even-numbered days received one treatment, and those admitted on odd-numbered days received the other. That assignment sounds random enough, but it failed. The number of patients admitted with the disease on even days gradually became larger than the number admitted on odd days. Why? What happened is that some of the physicians figured out the scheme and made it a point to admit their patients on days when the procedure they preferred was in use. A bias such as this makes it possible for the patients admitted on even and odd days to be quite different. You cannot rely on the results of a study that used nonrandom assignment.

"BLIND" EXPERIMENTS

In experiments, as in surveys, you must not bias your observations or treatments with your own opinions or preconceptions about which group or treatment should yield better results. Some events, of course, are not disputable, such as the fact that a rat has died. However, when making observations that are not as clear-cut, such as assessing the happiness of a person's marriage, it is all too easy to let unreliable judgment creep in — even though you are trying to be objective and "scientific."

Not only you as an experimenter but also your subjects (especially if they are humans) can influence the outcome of an experiment without even trying. An example of a biasing influence is the *placebo effect,* a well-known effect in medical research. A placebo (such as a brightly colored pill that has no real effect) and a pep talk from a sympathetic physician are enough to cure many ailments. In an experiment on alertness, for example, if students believe that the vitamin supplements they get with their math lessons are intended to make them less sleepy during class, they may actually feel more alert (or more drowsy if they have a bias against the experiment's success). In an experiment on anxiety, if the patients believe that the pill they are getting contains a drug with a powerful relaxing effect, they will feel more tranquil than if they believe that they are just getting breath mints.

The placebo effect can occur in many kinds of experiments, not just in medical research. To avoid the effect, you should prevent subjects from knowing which experimental group they are in, and you should not tell them anything about the expected results. Keep them "blind" as much as possible. Ethical considerations

require that they know about any risks and that they give "informed consent." However, you can still design the treatments to avoid biasing the results. For example, if one treatment requires a group of people to take pills, make sure that all of the other groups get pills too, even if they are just sugar.

The people who record the experimental results should also be unaware of the assignment of subjects to groups. They too should be "blind." Make sure they know exactly what to measure, such as weight without clothes, learning time to the nearest second, or anxiety on a particular scale. But avoid explaining more than they need to know. If you satisfy their curiosity by explaining what is going on while the study is in progress, you will never be sure whether they unconsciously affected the results. Explain the issues after the study is complete. You do not want anyone's prejudices to influence the measurements. Even if you are making the observations yourself, you can still keep yourself blind by not knowing which subject is in which experimental group. Have an assistant assign the subjects randomly to the various groups, leaving you pure and untainted.

Medical studies are often characterized as single blind or double blind. When only the subjects do not know which groups or treatments they have been assigned to, the experiment is called *single blind.* When both the experimenter and the subjects are kept unaware of the assignment, the study is called *double blind.* Double blind studies are the most reliable.

CONTROL GROUPS

If you are conducting a study to evaluate a new experimental method or treatment, make sure you include a group that does not receive the new treatment. This *control group* will provide you with measurements to which the results of the new treatment can be compared. If you are evaluating a new instructional method, for example, the appropriate control treatment may be the standard instructional method. If you are doing a medical experiment, the appropriate control treatment may be the standard medication or procedure for a particular ailment. If you are doing a study of a new component for a new sub-assembly, the control group may be an old design of that product.

Do not compare the new treatment's results just to historical information or commonly held beliefs. Experimenters may be tempted to do so, but then they run into a variety of problems. For example, a surgeon who is pioneering a new technique cannot simply compare the survival rates of patients who were given the new operation with those of patients from previous years. An engineer pioneering a new catalytic converter cannot afford to evaluate that new technology only by comparing it to past catalytic converters. Differences may occur for many reasons. Current patients may have been diagnosed earlier than previous patients, so they have a better chance of surviving. Another possibility is that the surgeon's skills may have improved with time, making the newer patients more likely to survive. In the case of the catalytic converter, it may be that the new one is "better" because it is positioned closer to the manifold or because it includes more precious metal.

All kinds of things may be different between groups that are treated at different times. You do not know — you cannot know — what all of these things are and

how they affect a study. To avoid this problem, make sure that a control group is part of your study's design, and do not rely on historical controls.

HOW SHOULD YOU PROCEED IF YOU WANT TO EXPLORE AN IDEA?

Here is what you should do when you want to design a study to explore an idea or question:

- You should carefully formulate your question and decide exactly what pieces of information are necessary to answer it.
- You must determine the population of interest and select a random sample of objects or people from the population.
- You must be sure that you do not unintentionally bias your sample by making it more likely that some members are included than others.
- You must collect your information in an objective fashion. The procedure for gathering the information must be objective and standardized. Questions must be unambiguous.
- If several different conditions are to be compared, you must ensure that the subjects are randomly allocated to the groups.
- You must prevent the subjects and investigators from allowing their personal prejudices to influence the outcome of the investigation.

SELECTED BIBLIOGRAPHY

Deming, W.E., *Some Theory of Sampling,* Dover Publications, New York, 1950.

Sudman, S. and Bradburn, N.M., *Asking Questions: A Practical Guide to Questionnaire Design*, Jossey-Bass, San Francisco, 1982.

Williams, B., *A Sampler on Sampling,* John Wiley & Sons, New York, 1978.

1 Designing and Using Forms for Studies

This chapter focuses on the need for order in running an experiment. It begins by addressing the need for having a specific form to collect the data and then proceeds to address some of the issues that every experimenter should be aware of when doing surveys and experiments.

OVERVIEW

Before you accumulate any data, you must design the questionnaire (for a survey) or have a very good idea about your operational definition and process (for an experiment), and you must have an appropriate form for collecting the data. It sounds both simple and silly to talk about these items, but they are fundamental to the integrity of the study.

For example, instead of asking directly for a person's age, you may want to request the actual month, day, and year of birth. Some people, especially as they grow older, have unpredictable systems for altering their ages at each birthday. Asking for the date of birth, instead of asking directly for age, increases the likelihood of an accurate answer. The computer will calculate the exact age later. Never ask interviewers or respondents to calculate anything themselves. Get the raw numbers, and let the machine do the arithmetic. If you are interested in the ratio of a person's weight to height, for example, the interviewer or respondent should record the weight and height, and you should leave the ratio to the computer. This procedure saves time and increases the accuracy of the results. Computers divide better than distracted people with calculators do.

CODING THE DATA

The answers to some questions are numbers. If you ask how much people weigh, how many cigarettes they smoke daily, or how many brothers and sisters they have, the answers will be numbers, and you can simply leave space on the questionnaire form to write in each one. You should leave enough room for the biggest number possible, even if you do not expect to get it. If you ask enough people how many children they have, for example, you will certainly find somebody who has ten or more, so you should leave sufficient room for a two-digit number.

When the answer to a question is not a number, you should try to figure out in advance what answers are possible. Respondents would then select among the alternatives. If you do not think of the possible responses before the survey, you will be in serious trouble. For a question about how people view life in general, if you let people supply their own alternatives, you may end up with as many different answers as there were people. How are you going to analyze "Kinda OK," "Could be worse," "Great, except for my job," "Today exciting, yesterday not"? You would spend hours deciding what to do with a few hundred of such answers. By forcing people to choose among specific alternatives (such as "Exciting," "Routine," "Dull"), you can get data that you can analyze.

What if you really are interested in the way people say things on their own? Sometimes you simply do not know in advance what people will say, and you want to allow them to say exactly what they please. You can certainly have interviewers write down exactly what the respondents say, word for word. Questions like these, which do not specify the possible responses, are known as *open-ended questions.* A computer will not be much help in analyzing the responses to open-ended questions directly, though, so you should either study the answers and assign codes to them before you enter your data or else ask the same questions again in a different way, with specified choices for responses.

For questions that require choosing among alternatives, think about all the possibilities. Make sure nothing falls through the cracks. Anticipate the unusual. For example, if you want to ask about housing, remember that not everyone lives in a house or an apartment. It is especially important to make provisions for responses such as "Don't know" or even "None of your business." Do not leave it up to an interviewer to decide what to do when somebody cannot or will not answer a difficult question. Anticipate these problems, and write clear instructions on the form in the places where such answers can occur. Whenever answers that do not fit into your coding scheme are possible, include an "Other" category, and leave space on the form for writing out the unusual answers. You may be able to do something with this written information later.

On any survey form, all acceptable answers to a question should be listed. Each answer has a code with it — a number that represents that answer. For the exciting-routine-dull life question, a code of 1 is circled for the answer "Exciting," 2 for "Routine," and 3 for "Dull." The code number 8 is reserved for the answer "Don't know." These numerical values are the *coding scheme* for the variable. For each respondent, one of these numbers will go into the computer to represent the answer to this question.

Coding schemes are arbitrary. A code of 3 could just as well have been assigned to Exciting and a code of 1 to Dull. What is important is that each possible response has a code that is different from the others. For example, you would not code the states of the union by their first letters because the first letters for the names of many states are the same.

Coding is an excellent way to do analysis with variable data, whether those data come from engineering, marketing, personnel, or any other place to which the study is directed.

TIPS ON FORM DESIGN

Here are a few hints about things you can do when designing your form that will make your life a lot easier when it comes time to analyze the data. Some of these hints pertain to the arrangement of the form, and some are about coding.

To design a form well, you should:

- *Split up complicated questions.* Some questions are best asked in parts.
- *Record numbers when you can.* Record information in as much detail as possible, using actual numbers. Later, you can use the computer to create categories such as small, medium, and large based on the exact numbers. If you have recorded only the categories, you will not be able to try different grouping schemes or to analyze the data in more detail. (Exceptions to this rule exist. For example, if you are a marketer, you may need to ask people about their incomes. The questions that you ask should not focus on the precise income. That is because income is a sensitive matter for many people, and they may refuse to give their income as an exact number of dollars; or they may not know their yearly income to the nearest dollar. If you are sensitive to these types of questions ahead of time you can soften the question by letting the interviewer hand a card to the respondent with preprinted income categories. That way, the respondent never needs to give an exact figure. Analyzing income categories may be harder than analyzing exact incomes — but the situation would be even worse if people refused to answer at all or if they gave false answers.)
- *Use a numeric coding scheme.* When items require coding, assign numbers instead of letters to the responses. Numbers simplify both data entry and data analysis. For example, coding "comfortable ride" as 1 and "uncomfortable ride" as 2 is simpler than the actual adjective used in the study.
- *Put an identification number on the form.* You can use the identification number to locate forms that you later find to have errors or unusual values. Even if you are running a confidential survey, and each form is not linked to a particular person, put an identification number on the respondent's form *before* you enter the data into the computer. Then enter the identification number with the rest of the data. This number links the paper form and the computer record.
- *Make sure the data can be entered into a computer directly from the form.* This saves time and helps to minimize errors.

COLLECTING THE DATA

In many studies, much of the work comes when it is time to collect the data. Remember that the interviewers or experimenters must carry out your well-designed survey or experiment exactly as you specified. In a survey, they must ask the questions in a standardized way, without leading the respondents, and they must use your coding schemes by entering the proper types of information in the right places. In

an experiment, the experimenters must make sure that all the variables are accounted for on their particular sequence and level. It is important to note any irregularities for later analysis. Unless the interviewers (or you) gather the data well, all the work you did to prepare the study and any work you do to analyze it will be for naught.

WHAT COMES NEXT?

When you have the completed forms from your study, you need to enter the data into a computer so that the computer programs can carry out your instructions for data analysis. Each item of data entry is a "case" or "run" of the experiment. When you have entered all the data, make sure that all the data (1) is in and (2) is saved appropriately for future analysis. The mechanics of entering data into a disk file on your personal computer are pretty simple. If you are planning to analyze the data, however, you do not want to make mistakes at this point. You cannot get good analyses with bad data.

It is easy to make mistakes when you type in your data. Here are some tips that will make it easier for you to find and fix these mistakes:

- Line up the data for each case so that the same types of information always appear at the same positions on the line. If you have room, leave a blank space between the pieces of information or between groups of them so that they are easier to read. In other words, use *fixed format.*
- Start each case on a new line. If you have too much data for a case to fit on a single line, go ahead and use more than one line per case, but never put more than one case on a line. (You cannot, if you use fixed format.)
- Put the case identification number at the beginning of each line of the file. If you are using more than one line per case, repeat it on each line. Also number each line *within* the case.
- Save your data file frequently as you enter the data. That way, you will have a permanent copy of most of your data even if something goes wrong while you are working.
- Make a backup copy (an extra copy) of the data to use if your original data file is somehow lost or destroyed. Since disks can be damaged, put the backup copy on a different physical disk. If you are using a floppy-based personal computer, put the original and the backup on different floppies. If your computer has a hard disk, you can leave the original on the hard disk and put the backup on a floppy. Be sure to label the backup floppy diskettes clearly and put them in a safe place.

MODIFYING THE DATA

After you have entered the data into the file, a time may come when you need to modify your data before analyzing them. You can change the values of individual variables or perform calculations to create new variables out of existing ones. If you do this, you may need to go back and supply new labeling information to describe

the modified data. The process depends on the software. If you have done anything to change the data — such as revising the coding scheme of a variable or simply reading a raw data file into your original format — you should save a copy of the data in their existing form.

ANALYZING THE DATA

Any time after you have identified the data set, you can enter a command to analyze it. The command will depend on the objective of the project and the need for a particular technique. Depending on the software that is used, many analysis commands may be available.

PRINTING THE RESULTS

The results of your analysis are (1) displayed on your computer screen of or (2) printed in a hard copy format.

MISSING DATA

For each variable in a study, a special code indicates when information is missing. Data may be missing for legitimate reasons. Nevertheless, they have to be identified appropriately. Using computer software, this has become a very easy task. Depending on the software, the coding may take different formats in both identification and processing. (Usually, these codes are identified with the numbers 9; 99; 999 or (.)).

If you have used a specific code to identify missing data, the computer will usually treat the data as missing because you told it to, by entering a MISSING VALUE command. Data that are missing for this reason are called *user-missing,* because you (the user) specified them as missing.

However, sometimes the particular software must treat data as missing regardless of whether you tell it to or not. Perhaps a case in your data file is simply missing some variables. Perhaps somebody's fingers slipped on the keyboard and entered a response as "YP" instead of "60." When things like this happen, the statistical software assigns a special value called the *system-missing* value. Statistics are never computed with system-missing values because they are not proper values at all.

What good are missing values? Why do we have to fool with them? Most of the time, you cannot really do anything with missing values, but you do not want to throw away the whole cases they came from. Other variables in those cases may have perfectly good values. And you may change your mind about user-missing values. People who do not know for whom they will vote are sometimes useless for your analysis, but sometimes they are the most interesting people of all.

Nobody wants missing values in their data, but they always turn up. One of the first things you should do with a new data file is to get a general idea of how many missing values there are and why. You can do that with frequency tables, as will be explained in the next chapter.

2 Counting Frequencies

This chapter begins to explain how data can be analyzed using the basic statistical tools. The counting of frequencies should be the first step in data analysis for two reasons: (1) it enables you to screen the data for any irregularities, and (2) it allows you to group the data in percentages or some graphical format.

OVERVIEW

The data have been collected and entered into the computer, and now you are about to begin your analysis. Hopefully, this analysis will generate all the answers that you are seeking to explain.

Most experimenters begin their analysis by double-checking their data. Double-checking involves the actual counting of all entries by variable or anything else that has been coded. Fortunately, since you are analyzing your study with a statistical software package, you do not have to actually count anything. You can tell the computer to do it for you. Use the appropriate commands, and the answer will be in front of you before you know it. The FREQUENCIES command counts the number of times each of the codes occurs. You supply the names of the variables for which you want counts; everything else is done for you.

INTERPRETING A FREQUENCY TABLE

When you run a job with the FREQUENCIES command, you get back something called a *frequency table.* This is simply a table that tells you how frequently each of the responses occurs. Generally, the table will provide you with the file name, the variable, and the details of the frequency command, which include: value, actual frequency, percent, valid percent, cumulative percent, missing, and totals.

If you find codes in the frequency table that are not supposed to occur in the data, you need to go back and correct or at least check them.

VALID PERCENTAGES

Sometimes you want to compute percentages using only cases with real responses. For example, suppose you have asked 100 people whether life is exciting or routine, and 25 said that it is exciting, 25 said that it is routine, and 50 told you to bug off.

It would be a bit misleading, though it is correct, to state that 25% of those people think that life is exciting. A naive reader or listener would probably assume that the other 75% of the people find life unexciting. That is not really true, since the remaining 75% include people who declined to answer as well as those who find life routine. You can describe the results better by saying that half of the people who answered the question find life exciting, and half find life routine. You should also mention that half of the people in your sample refused to answer the question. You can find the percentages based only on cases with real answers (so-called *valid cases*) in the column labeled VALID PERCENT in the frequency table.

BAR CHARTS

Usually the frequency command has a column, labeled CUM PERCENT, which is also valuable. To transform your frequency table into a picture, you merely add a slash and the word BARCHART to your FREQUENCIES command.

When you execute this command, the computer produces a type of display that is called a *bar chart* because each line in the frequency table is turned into a bar. The length of the bar depends on the number of cases. (The actual frequency is given beside the bar.) At a glance, you can tell how often each of the responses was selected. You can also see whether one of the responses was an overwhelming favorite, and which responses are about equally likely.

Since computer screens and printers have a limited ability to show detail, responses that have similar frequencies may end up with bars of equal length even though the actual frequency counts are slightly different. This does not really matter. The point of a bar chart is to provide a visual summary of the data, and such minor distortions do not change the overall impression. If you want precision, look at the numbers, not the chart.

CUMULATIVE PERCENTAGES

Yet another statistic presented as an output of the frequency command is the CUM PERCENT. The cumulative percentage for a response is the sum of the valid percentages for that response plus all responses that precede it in a frequency table.

LEVELS OF MEASUREMENT

Depending upon the type of variable that you have used, the numbers in your data table may have different meanings.

Categories

If you think about the numbers used to code some of variables in many experiments and surveys, you will realize that the particular number assigned to a category conveys no *numerical* information. The codes just represent the categories.

Ordered categories

Sometimes the order of categories is significant. Think about the exciting-routine-dull variable. The responses to the question can be arranged in a meaningful order. If we arrange them in terms of decreasing excitement, then the response "Exciting" comes first, followed by the response "Routine," and finally the response "Dull." Of course, we could have arranged the responses in the other order as well (from low excitement to high excitement). In both instances, the response "Routine" falls between the other two. There is no *order* to those categories, but it does mean something that Routine is between Exciting and Dull.

Numbers

Although the codes assigned to the exciting-routine-dull variable are ordered from high to low, they convey only order; they have no other numerical meaning. Someone who was bored with life (code 3) did not differ by two "excitement units" from someone who found life exciting (code 1). Subtracting or dividing the codes makes no sense.

On the other hand, let us say that "education" is one of the variables under study. This variable is different. The numerical code assigned to each category is not merely a code. It is the highest grade completed. It is an actual number, and we can treat it as such. For example, someone with 8 years of education has twice the number of years of education as someone with 4 years. Someone with 16 years of education has 4 more years than someone with 12 years. We can add, subtract, and divide the codes and understand the results.

NOMINAL, ORDINAL, INTERVAL, AND RATIO

Variables can be classified into different groups based on how they are measured. Machine number, cost, and weight are all different types of variables. Machine number is called a *nominal* variable because the numerical codes assigned to the possible responses convey no information. They are merely labels or names. (That is why the level of measurement is called nominal — from the Latin word for "name.") Codes assigned to possible responses merely identify the response. The actual code number means nothing.

If the possible responses can be arranged in order, as with the exciting-routine-dull variable, the variable is called *ordinal***:** its codes have an order, nothing more. (The word *ordinal* comes from — you guessed it — a Latin word meaning "order.") Variables such as dollars, job satisfaction, condition of health, and happiness with one's social life, all of which are usually measured on a scale going from much to little, are ordinal variables. The numbers assigned to the responses allow you to put the responses in order, but the actual distances between the numeric codes mean nothing.

Temperature can be measured and recorded on a scale that is much more precise than job satisfaction. The interval or distance between values is meaningful everywhere on the scale. The difference between 100 degrees Fahrenheit and 101 degrees

Fahrenheit is the same as the difference between 102 degrees and 103 degrees. Since temperature measured on the Fahrenheit scale does not have a true zero, however, you cannot say that an 80-degree day is twice as hot as a 40-degree day. A temperature of zero does not mean that there is no heat. The zero point is determined by convention. (If you insist, I will admit that temperatures do have an absolute zero point; but that has very little to do with the measurement of body or environmental temperatures.) Thus temperature can be called an *interval* variable.

The last type of measurement scale is called a *ratio* scale. The only difference between a ratio scale and an interval scale is that the ratio scale has an absolute zero. Zero means *zero.* It is not just an arbitrary point on the scale that somebody happened to label with zero. Height, weight, distance, age, and education can all be measured on a ratio scale. Zero education means no education at all, and zero weight means no weight at all. On a ratio scale, the proportions or ratios between items are meaningful. A 200-pound person is twice as heavy as a 100-pound person. A 1000-meter race is twice as long as a 500-meter race. I suppose I need not tell you what language the words *interval* and *ratio* come from.

Why all the fuss? Why have we spent all this time describing these "levels of measurement?" The reason is straightforward — the way in which you analyze your data depends on how you have measured it. Certain analyses make sense with certain types of data. Even something as simple as interpreting cumulative percentages requires you to know what scale your data are measured on. For example, cumulative percentages do not make much sense for variables measured on a nominal scale. So, depending on the level of measurement, the appropriate technique, test, and analysis must be selected. Otherwise, the results will be meaningless.

3 Summarizing Data

In the previous chapter, you saw that a frequency table is a convenient way of looking at the responses to a question. Frequency tables are easy to read, and they provide complete and detailed information. Sometimes, though, they provide too much information. To bring the information into focus and make the data come alive when communicating the findings of your study, you need to group and summarize your data. This chapter tells you how to do that.

DESCRIPTIVE STATISTICS

Think about a variable such as age, height, temperature, pressure, or weight. When you measure these variables, a lot of different responses are possible. The more finely you measure the variables, the larger the number of possible responses. For example, if you record height only to the nearest foot, the number of different heights is fairly limited. However, if you measure height to the nearest millimeter, it is possible that everyone in your sample might have a different measurement. What would happen if you made a frequency table for such a variable? You would probably end up with an enormous table with a lot of different values. Most of the codes in the table would show only a single case with that particular value. In fact, if every case had a different value, you would end up with nothing more than a list of all the responses. That kind of frequency table does not do much for you. You need some way to summarize the data further.

Nominal Variables

The way to get further summaries depends on how the variable is measured. If you have a frequency table that shows the number of people who were born in each of 400 cities, you may not be able to summarize further at all. Since the name or ID number of the city is a nominal variable, you cannot group the cities into larger categories without further information, such as what state they are in. In other words, you cannot really summarize a name effectively. All you can do with the name or ID number is count the number of people for each one.

If you are going to report the results to an audience that has a short attention span, you might organize the frequency table so that it goes from the most frequent

city to the least frequent. Then you could mention only the "top ten" cities. The "top" city has a statistical term you can use to describe it — the ***mode***.

Nominal variables with many categories simply do not lend themselves to summarization by computer. If you need to summarize, you have to rearrange the coding system. For example, you can combine cities in the same state, or you can group them by population. You can then make frequency tables based on the new, more compact classification.

Ordinal Variables

It is easier to summarize an ordinal variable than a nominal variable. If you make a frequency table and decide that you have too many categories (Extremely exciting, Greatly exciting, Moderately exciting, Mildly exciting, Slightly exciting, Almost but not quite exciting...), you can combine adjacent categories. One way to do this is to convert all the different codes that stand for varying degrees of excitement to a single code: Exciting. Similarly, you can combine the various codes for Routine and for Dull. The frequency table for the less elaborate coding scheme will be easier to read and probably just as informative.

A variable with ordered categories also gives you more choice in descriptive statistics. You can report the mode (the category that has the largest number of cases) for an ordinal variable, as you can for a nominal variable. The mode, remember, tells you which response occurred most frequently. In addition, another value, often more descriptive, can be computed for an ordinal variable. It is called the *median.* The median is the "middle" value — the value that divides the observations into equal halves. Notice that you cannot have a middle value unless it makes sense to put the values in order. That is why the median is a useful statistic for ordinal variables but not for nominal variables.

If you ask five people to rate the president's performance on a scale of 1 to 5, and you get the answers 1, 1, 3, 4, 5, the median answer is 3. The value 3 divides the five responses into equal halves, when they are placed in order like this. The median is the middle observation when the values are ordered from smallest to largest. The median provides you with some idea of what a typical response is.

What if you have an even number of observations? There may not be a single median, since two numbers are in the middle. With the numbers 1, 2, 3, 4, the numbers 2 and 3 are equally in the middle. If this happens, you can still calculate the median. Identify those two middle numbers, and figure out what number would be in the middle of them in this way:

- Add the two middle numbers together.
- Take half of their sum.

In this example, you would add the middle numbers 2 and 3 to get 5. Half of 5, or 2.5, is then the median of 1, 2, 3, 4. Statistical software programs do this for you automatically.

Interval or Ratio Variables

If your variable is measured on an interval or ratio scale, you can summarize it in many different ways. Your options for more powerful analyses are much greater than they were with the other types of data. You could still make a frequency table, but it would probably be unwieldy and not particularly informative. Transforming the frequency table into a bar chart probably would not help, since the chart would have as many bars as there are different values. It would be more useful to make another frequency table in which each line represents not a single response but several ones. In other words, you should group the responses into categories. This approach turns out to be more manageable and more expedient. You can then use a modification of the bar chart, called a *histogram*, to display the number of cases occurring in each of the categories. With most software programs, you can create both a frequency table and a histogram by identifying such a request under the FREQUENCY command. A histogram gives you information about the total count and the midpoint, shape, and spread of the distribution. The *minimum, maximum,* and *increment* specifications are optional. They determine the lowest and highest values shown, as well as the size of the interval. This is important because, as we are going to see later, you can use this information to determine the capability of a process just by utilizing the histogram.

The number of intervals you should use in a histogram depends on the data. If the intervals are very wide, you may not be able to see important differences. On the other hand, if they are too narrow, you may have more detail than you want to see. A good practice is to do several histograms and see which one summarizes the data most clearly. A rule of thumb for generating the groupings of the data is to calculate the square root of the number of observations. For example, if you have 100 observations, groups of ten would be appropriate because $\sqrt{100} = 10$.

Differences between Bar Charts and Histograms

A histogram looks pretty much like a bar chart. Only two real differences exist:

1. In a bar chart, each bar represents a single code, while in a histogram the bars often represent the combined frequencies of several codes.
2. Bar charts and histograms treat codes with no cases (frequencies of zero) in different ways.

To make a bar chart, you do not have to assume anything about what the codes actually mean. If you are using the codes from 1 to 3, and no cases have a value of 3, there simply is no bar for that code. Since you can use whatever codes you want for nominal and ordinal variables, the computer cannot tell what codes were possible but did not occur. If a value has no cases, no line appears for it in a frequency table, and no bar appears in a bar chart.

On the other hand, if the variables are measured on an interval or ratio scale, you do want to know when some of the values do not occur in your data. When this happens, a histogram leaves a space for them with no bar. If no people were in a particular sample or category, the histogram would have room for that category, but

the bar would have shrunk to zero. Conversely, if you made a bar chart of the ages, on the other hand, no space would be left for the particular category. The "holes" in the histogram tell you that some possible values did not occur at all. That makes it easier for you to see what the real distribution of values looks like. It is essential to know about the "holes" when you have an interval or ratio variable.

Uses of Histograms

Histograms are useful whenever:

1. A variable has many different values.
2. It is reasonable to group adjacent values.

Never use a histogram to summarize a nominal variable. By looking at a histogram, you can see the shape of the distribution and therefore you can learn:

1. How often the different values occur.
2. How much spread or variability exists among the values.
3. Which values are most typical of the data.

These things are important, first of all, because they tell you a lot about your data. Also, some of the statistical procedures that we will be using later do not work properly unless the data come from particular types (shapes) of distributions.

Different Types of Distributions

A variable such as age can have many different types of distributions, depending on the population you study. If you are studying children who are entering the first grade, you will discover that their ages are fairly similar. If you made a histogram of the ages, you would most likely end up with two long bars, one for 5-year-olds and one for 6-year-olds, with a few short fragments for 7- or 8-year-olds.

On the other hand, if you study college freshmen, you will find that the distribution of the ages spreads out more. Although the majority of college freshmen are either 18 or 19, there are always a few younger students who skipped grades, and surprisingly often there is an octogenarian catching up on what he or she missed. You find people of many different ages in this sample. Some values are more likely than others, but many different ones occur.

Finally, if you are studying the entire U.S. population, your sample includes people of all ages. The histogram of their age distribution would not look anything like that of the first graders or of the college freshmen.

It is fairly obvious that the distribution of the variable "age" will vary depending upon the group of people under study. In other situations, the fact that distributions vary in different populations may be less obvious — but it is no less important. For example, consider the ambient temperature at which a particular product must operate. If you test the product in Arizona, Louisiana, Florida, and Alaska, you will get different responses at each site. If you collectively take responses from Mexico,

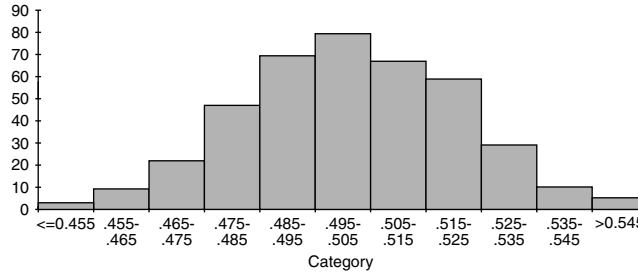


FIGURE 3.1 A typical histogram showing normality.

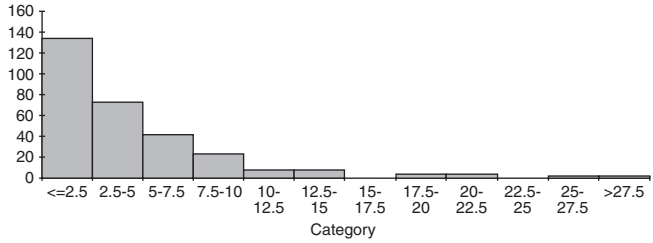


FIGURE 3.2 A typical histogram showing a positive skew distribution.

the United States, and Canada, the distribution will look completely different from the distribution you would have obtained at one particular location. Figures 3.1 through 3.4 show histograms with different distributions.

More Descriptive Statistics

Often, you want to summarize data even further than a histogram allows. You would like to be able to report some numbers that describe the distributions more precisely. What sorts of descriptions might these be? The mode — the most frequently occurring value — is the simplest way to represent "typical." For a nominal variable, it is the only thing we can use. The median — the middle value when values are

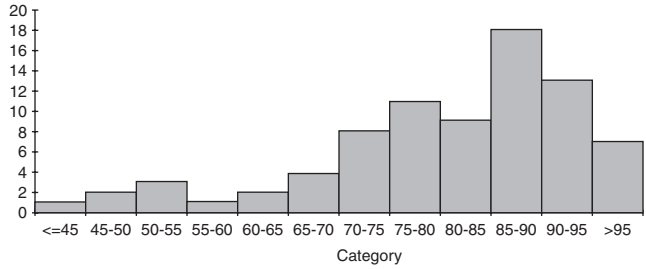


FIGURE 3.3 A typical histogram showing a negative skew distribution.

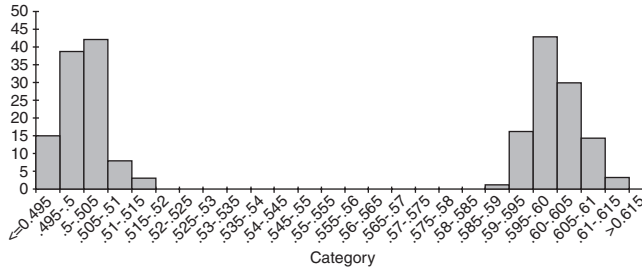


FIGURE 3.4 A typical histogram showing a bimodal distribution.

arranged from smallest to largest — is another way of representing "typical." Of course, you can only make sense of the median for a variable that is measured at least on an ordinal scale.

In addition to the mode and the median, other handy statistics can be used to describe your data.

Other Percentiles

The median is the value that splits the sample into two equal parts. Sometimes, though, it is useful to look at values that split up the cases in other ways. What is the value that cuts off the bottom quarter of the cases or the top quarter? These values are called *percentiles* because they tell the percentages of cases above and below them. The median is the 50th percentile, since 50% of the cases have larger values and 50% have smaller values. The 25th percentile is the value that splits the cases so that one quarter of them have values below it. (It follows that 75% of the cases exceed the 25th percentile.) If you have made a frequency table, you can locate percentiles in the cumulative frequencies column. However, you can go about it in a simpler way. By identifying the subcommand percentiles in the frequency command, you will get the percentiles.

The Average or Arithmetic Mean

For interval and ratio variables, the arithmetic mean or average is usually a better measure of central tendency than either the mode or the median. It is simple to calculate. Just add up all of the values and divide the sum by the number of cases. When you are using a statistical software package, you can get this information just by requesting it.

When you calculate (or have your computer calculate) the mean, median, and mode of a set of data, you will often notice that the three values — each of which represents the "typical" value of the distribution — are different. Why are all of these "typical" values different? There is no reason for the numbers to be identical since they all define "typical" in different ways. The mode is the value that occurs most often; the median is the middle value when the numbers are arranged from smallest to largest; and the mean is the familiar "average" value. Which of these is the best measure of "typicalness" (more formally called *central tendency*)?

Mean, Median, or Mode?

Usually the mode is a poor measure of central tendency for an interval or a ratio variable. Though it satisfies one of the definitions of "typical," it ignores much available information about the data.

Although the median is a good measure of central tendency, it ignores a lot of the information that you have collected about a variable measured on an interval or ratio scale. For example, the median of the five ages 28, 29, 30, 31, and 32 is 30. The median for the five ages 28, 29, 30, 98, and 99 is also 30. The actual values of ages above and below the median are ignored. The median is 30 regardless of whether everyone's age is close to 30 or whether the values vary quite a bit.

When should you report the median, and when should you report the mean? If a variable is measured on an ordinal scale, the median is the statistic of choice. If a scale does not have intervals of equal length, it does not make sense to compute a mean. For a variable measured on an interval scale, the mean and the median are both useful numbers to report. The mean makes maximum use of the data since all of the values are actually used in computing it. (Remember, you add up all the numbers, then divide by the number of numbers.) In some situations, however, the mean may not really represent the data well.

Suppose you ask five people how many parking tickets they have received in the last year, and you get the following replies: 2, 5, 6, 7, 90. The mean number of tickets for this sample is 22. (Verify this: the sum is 110, and 110 divided by five is 22.) This statistic does not describe the data well. The person who hardly feeds a meter is making the other people in the sample look more delinquent than they really are. The median, six, describes the data better.

Whenever some cases have values much larger or smaller than the others, the mean may not be a good measure of central tendency. It is unduly influenced by extreme values (called *outliers*). In this situation, you should report the median and mention that some of the cases had extremely large or small values. For example, you could say, "The median number of tickets for the sample is six. Eighty percent had seven or fewer tickets a year. One person reported 90 tickets."

HOW MUCH DO THE VALUES DIFFER?

Measures of central tendency provide information only about "typical" values. They tell you nothing about how much the values vary within the sample. Suppose you examine the transmissions of 20 vehicles: ten trucks and ten sedans. The trucks were used to haul heavy items, whereas the sedans were used for pleasure driving. You are interested in measuring the problems with each vehicle during three months in service. The data are shown in Table 3.1.

The average number of problems for the two groups of vehicles is the same — five per week. However, the distributions of values differ. All of the trucks had consistent numbers of problems — between four and six per week. The numbers do not vary much from week to week. The sedans, on the other hand, differ from each other much more. Some seem to have no problems early on but then as time goes on their problems increase.

TABLE 3.1
Vehicle Problems per Week

Trucks		Sedans	
Problems per Week	Difference	Problems per Week	Difference
4	–1	0	–5
4	–1	0	–5
5	0	0	–5
5	0	0	–5
5	0	0	–5
5	0	3	–2
5	0	10	5
5	0	10	5
6	1	12	7
6	1	15	10
Sum = 50	Sum = 0	Sum = 50	Sum = 0
Mean = 5	Mean = 0	Mean = 5	Mean = 0

How can you measure this variability? One of the more obvious ways is to report the smallest and largest values in each of the samples. The minimum number of problems for the trucks is four and the maximum is six. In the sedans, the minimum number of problems is zero and the maximum is 15.

The distance between the largest and smallest values is called the *range***.** For the trucks, the range is 2, while for the sedans, it is 15. That is quite a difference. By comparing the ranges of the two samples, you can tell that the number of problems per vehicle in the sedan sample differed more from each other than those in the truck sample.

The range is not a particularly good measure of variability, though. It depends only on the smallest and largest numbers and pays no attention to the distribution of the numbers in between. Nevertheless, it is the best measure of variability available for variables measured on an ordinal scale. For a variable measured on an interval or ratio scale, however, you can compute some better measures.

The Variance

For each case, you can compute how much it varies from the mean of all the cases. Just subtract the overall mean from the case's value. For the first case in Table 3.1, the difference is:

$$4 \text{ (the case's value)} - 5 \text{ (the mean)} = -1$$

This indicates that this particular truck had one fewer problem than the average. Table 3.1 shows the differences for each case. From the table, you can see that the differences are much smaller for the trucks' performance than for the sedans.

How can you use these differences to measure variability? The simplest tactic that comes to mind is just to add up the differences and compute a mean difference for each

group. Like many seemingly good ideas, this one has a flaw. The sum of the differences from the mean is always zero. Some of the differences are positive, and some are negative, so when you add up all of the positive and negative numbers the result is always zero. You need a better way to assemble all of the differences from the mean.

You can do this in several ways. For example, you could treat all the differences as if they were positive and compute a mean difference for them. It turns out, though, that a better way by far is to

1. Square the differences.
2. Add them up.
3. Divide the sum by the number of cases minus one.

This measure is called the *variance*.

Why divide by the number of cases minus one instead of the number of cases? You are working with a sample taken from a larger population, and you are trying to describe how much the responses vary from the mean of the entire population. However, since you do not know the population mean, you have to use the sample mean in your calculation — and using the sample mean makes the sample seem less variable than it really is. When you divide by the number of cases minus one, you compensate for the smaller variability that you observe in the sample. Later on, we are going to define this as degrees of freedom (df).

Large values for the variance tell you that the values are quite spread out. Small values indicate that the responses are pretty similar. In fact, a value of zero means that all of the values are exactly equal. For the set of data shown in Table 3.1, the variance for the trucks is .44, while that for the sedans is 36.44. This supports our observation that sedans vary more.

The Standard Deviation

Since you calculate the variance by squaring differences from the mean, it is expressed in a unit of measurement such as squared hours, squared children, or something similar. To express the variability in the same unit as the observations, you take the square root of the variance. This is called the *standard deviation*. The standard deviation is expressed in the same units as the original data. For the trucks in Table 3.1, the standard deviation is the square root of .44, or .66. For the sedans, it is the square root of 36.44, or 6.04.

With any statistical package, it is easy to calculate various measures of central tendency and variability. Under the FREQUENCY command, you specify the statistics kurtosis, mean, median, mode, standard error of the mean, skewness, sum, standard error of skewness, minimum, maximum, range, variance and standard deviation and the computer will do the rest.

Frequency Tables Versus Cross-Classification Tables

Up to this point, we have been dealing primarily with frequency tables. We can obtain additional types of information about our data, though, if we use cross-classification tables.

TABLE 3.2
Cross-Tabulation Table

Ride	Count	Male 1	Female 2	Row Total
Comfortable	1	300	384	684
		50.3	44.4	46.8
Normal	2	267	437	704
		44.8	50.5	48.2
Rough	3	29	44	73
		4.9	5.1	5.0
Column Total		596	865	1461
		40.8	59.2	100.0
Missing Observations = 12				

Fundamentally, the difference between frequency tables and cross-classification tables is that in the first, we deal with variables one at a time while in the second, we have the capability to do multiple comparisons. To do the comparisons, however, we must know the variables and their individual groupings.

For example, consider a study in which we have asked people to evaluate the ride of a new vehicle, classifying it as "comfortable," "normal," or "rough." We want to know how many of the 684 people who found the ride comfortable were men and how many were women. How do we find out? No problem. Just use the CROSSTABS command.

This command creates a table (such as Table 3.2) showing the categories of the two variables and compares the variables with each other. Notice that the numbers appear in *cells*, and they are arranged in rows and columns. Labels at the left and the top of the table describe what is in each of the rows and columns. To the right and at the bottom of the table are totals — often called *marginal totals* because they are in the table's margins.

Because the categories of the two variables are "crossed" with each other, this kind of table is called a cross-classification table or simply a *cross-tabulation***.** A cross-classification table shows a cell for every combination of categories of the two variables. Inside the cell is a number showing how many people gave that combination of responses. (For our example we used a 2×3 cross-tabulation table. However, the process is the same and just as easy with a computer even with more than two variables.) The table is a very efficient way to present a lot of numbers. When you get used to it, it is quite easy to read. Let us look at what is in the cells.

The number in the first cell of the table, 300, tells you that 300 males found the ride comfortable. The next number is in the column labeled Female, and it tells you that 384 females found the ride comfortable. The sum of these two numbers – 684 – is shown in the last column.

Each of the cells also contains a second number. This number is the percentage. For example: 50.3% of the men reported a comfortable ride, but only 44.4% of the women reported the same thing. These percentages are just the opposite of what the

counts show. It is easy to mislead yourself if you compare just the counts in the cells of a cross-tabulation table. It is better to turn the counts into percentages in order to eliminate the differences that show up when you have more people in one group than in another.

What do these numbers tell you? They tell you how likely it was that a person who considered the ride comfortable was male or a female. But you are probably interested in knowing how likely it was that a male or female found the ride comfortable, and the row percentages do not tell you that. It is usually true in a cross-tabulation table that either row percentages or column percentages answer your question. Deciding to use one or the other is often based on which variable you consider dependent, and which one you consider independent. This is very important, and is easy to remember which is which. Here is what you need to remember:

- The dependent variable depends on the other one.
- The independent variable does not depend on the other one; it goes its own way, independently.

In summary, here are some key points about how you study the relationship between responses to two or more questions which have a small number of possible answers:

- A cross-tabulation shows the numbers of cases that have particular combinations of responses to two or more questions.
- The number of cases in each cell of a cross-tabulation can be expressed as the percentage of all cases in that row (the row percentage) or the percentage of all cases in that column (the column percentage).
- The variable that is thought to influence the values of another variable is called the *independent* variable.
- The variable that is influenced is called the *dependent* variable.
- If there is an independent variable, percentages should be calculated so that they sum to 100 for each category of the independent variable.
- When you have more than two variables, you can make separate cross-tabulations for each of the combinations of variables.

Means

So far we have looked at the relation between variables and groupings of variables. By doing so, however, we have ignored some of the available information. All cases with values in the same range have been treated as the same.

We can look at the relation between ride and male/female in another way that still produces compact tables but is based on each person's actual preference. What we can do is to compute means. That is, we can compute the individual mean of the participants as well as the group. This can be accomplished by a simple command of "means" in any statistical package, provided that you specify the variables that you want to work with.

Means from Samples

So far we have tried to answer questions such as: "What percentage of the sample thinks that the ride is comfortable?" or "What is the average age of the people who said the ride is rough?" The emphasis was always on reporting just the results of the study. We looked at the data and described the sample. Nothing more.

Now, we will begin to look at the problems we face when drawing conclusions about a whole population on the basis of what is observed in a sample.

In our sample, the men were more likely than the women to find the ride more comfortable. No doubt about it: 50% of the men but only 44% of the women called the ride comfortable. Unless an error was made somewhere in entering the data into the file, the results are crisp and clear. We can speak about the sample with confidence. We know, or can figure out, anything we want to about the sample — assuming that we asked the right questions and had the data entered correctly into the file.

But talking about a sample is usually not enough. We do not want conclusions about the 1473 people in the study; we want conclusions about the population that this sample represents. We want to be able to say such things as: "A comfortable ride means more to American men than it does to American women." Based on the results in the sample, we want to speak about the population from which the sample was selected.

That may not seem like a big deal. Why not just assume that whatever is true for the sample is also true for the population? If the men in the sample found the comfortable ride more appealing than the women did, why not claim that the same must be true in the population? Let us just conclude that American men are more enthusiastic about comfortable ride than American women. That would certainly be simple. But would it always be correct?

Problems in Generalizing

Suppose you had a sample of two men and two women, and you found that one of the men but neither of the women was concerned about the ride. Would you be willing to draw the conclusion that, in general, men are more excited by the ride than women are? The numbers, especially if you do not think about them, suggest the headline, "Amazing new research shows that half of all men but no women at all are excited by a vehicle's 'ride.'" It does not take much statistical know-how to find fault with this headline. Generalizing from a tiny sample of two men and two women to the whole U.S. population is laughable. If you sampled another two men and two women, you would probably get completely different results. But you could not generalize from those results, either. You cannot conclude much at all about the whole population from a sample of four people.

What if the sample were larger, say 200 men and 200 women? Conclusions from a study with this sample size would certainly be more believable than those from a four-person study. It is easier to believe that the results observed in the larger sample hold true for the population. But if you found that 50% of all the men and 49% of all the women in the larger sample said their vehicle ride was comfortable, would you be willing to conclude that in the population, men are more likely than women

to find the ride comfortable? What if the difference were larger, say 50% of the men compared to 40% of the women?

Sampling Variability

You get different results from different samples. Consequently, it takes some thought to sort out what you can reasonably say about the population, based on the results from a sample. If you and I each look at samples of 400 people from the same population, we are not going to get exactly the same answers when we each analyze our own data. Our samples will undoubtedly include different people, and our results will differ. With any luck, the results will be similar, but it is very unlikely that they will be identical to the last decimal place. Even if they are, they probably would not be the same values that we would obtain if we questioned the whole population.

How much the results from different samples vary from one to another depends not only on the size of the samples but also on how often the various responses occur in the population. (Statisticians call this the distribution of responses in the population.) If everybody in the United States plans to vote for the same candidate for President — say, the one you have been working for — any old sample will lead to the same answer to the question, "What percentage of the vote will my candidate receive?" The answers would not vary from person to person, and they would not vary from survey to survey. Any survey would tell you that 100% of the voters plan to vote for your candidate.

On the other hand, if only half of the voters plan to vote for your candidate, your samples would show more variability. One sample might show that 60% of the vote will go to your candidate, and another sample might show that your candidate will get 45% of the vote. If 1000 researchers took random samples of 400 voters each, they would obtain a lot of different percentages. Some would be close to the correct figure of 50%, while others would be higher or lower.

By now you should be wondering if the size of the sample has anything to do with the results of the study. The effect of sample size on any study is indeed important. For our discussion here, a basic fact to remember is that results from large samples do not vary as much as results from small samples do. You can test this on your own with a simple simulated study.

A Computer Model

We can use the computer to actually do what we have been talking about. With the proper instructions, it can set up a population in which half of the people say they will vote for your candidate, and half say they will not. We can instruct the computer to conduct a hypothetical survey by randomly selecting 400 cases from this population. Then we can tell the computer to calculate from this sample the percentage of the cases that endorse your candidate. We can have the computer repeat this kind of survey as many times as we want. Each time, it will select a new random sample of 400 hypothetical people and compute the percentage planning to vote for your candidate. (This is called a simulated survey.) This is very important for you to recognize because we are going to use this principle in Volume VI, titled *Design for Six Sigma*.

Other Statistics

Although we examined the percentage of people planning to vote for a candidate, we could have looked at some other characteristic, such as mean weight, mean number of pencils owned, or mean income. The procedure would have been the same. For each of the random samples from a population, we would have calculated the mean. Then we would have seen how much the mean values varied from sample to sample. The results would have been very similar to what we have seen, and the same basic rules would have applied. After all, percentages agreeing with a statement are equivalent to means.

How can a percentage be the same thing as a mean? For a variable that can have only two possible values (such as yes or no, agree or disagree, cured or not cured), you can code one of the responses as 0 and the other response as 1. If you add up the values for all of the cases, divide by the number of cases, and then multiply by 100, you will obtain the percentage of cases giving the response coded as 1.

Consider a simple example. You ask five people whether they approve of the president's performance. Three say they do, and two say they do not. If you code Approve as 1, you have the values 1, 1, 1, 0, 0. The mean of these values is 3/5 = .6. To get the percentage agreeing with the statement, just multiply the mean by 100. In this survey, 60% of the people approved of the president's performance.

At this point it is important to take a breather. We have introduced the word *statistic* quite a few times but without an official explanation. So, what is a statistic? A statistic is nothing more than some characteristic of a sample. The average height of the people in a sample is a statistic. So is the standard deviation or the variance of the heights. The term *statistic* is used only to describe sample values. The term *parameter* is used to describe characteristics of the population. If you could measure the height of all the people in the United States and calculate their average height, the result would be a parameter, since it would be the value for the population. Most of the time, population values, or parameters, are not known. You must estimate them based on statistics calculated from samples.

Here is what you can you say about the mean of a population, based on the results observed in a sample:

- When you take a sample from a population and compute the sample mean, it will not be identical to the mean you would have obtained if you had observed the entire population.
- Different samples result in different means.
- The distribution of all possible values of the mean, for samples of a particular size, is called the *sampling distribution of the mean.*
- The variability of the distribution of sample means depends on how large your sample is and on how much variability exists in the population from which the samples are taken.
- As the size of the sample increases, the variability of the sample means decreases.
- As variability in a population increases, so does the variability of the sample means.

Describing Data Sets with Boxplots

The final tool to be discussed in this chapter is the boxplot, a very useful graphical method for summarizing data. Boxplots can be used in two ways: either to describe a single variable in a data set or to compare two (or more) variables. The keys to understanding a boxplot are the following:

- The right and left of the box are at the third and first quartiles. Therefore, the length of the box equals the interquartile range (IQR), and the box represents the middle 50% of the observations. The height of the box has no significance.
- The vertical line inside the box indicates the location of the median. The point inside the box indicates the location of the mean.
- Horizontal lines are drawn from each side of the box. They extend to the most extreme observations that are no farther than 1.5 IQRs from the box. They are useful for indicating variability and skewness.
- Observations farther than 1.5 IQRs from the box are shown as individual points. If they are between 1.5 IQRs and 3 IQRs from the box, they are called *mild outliers* and are hollow. Otherwise, they are called *extreme outliers* and are solid.

Boxplots are probably most useful for comparing two populations graphically. As for the terminology and the conventional interpretation of the plot, we owe it all to the statistician John Tukey.

4 Working with the Normal Distribution

This chapter will explain how to use distributions of the mean to calculate confidence intervals for the population mean. It will also show how to use them to evaluate hypotheses about a population mean. The rest of the book will use ideas from this chapter to evaluate hypotheses of many kinds.

OVERVIEW

Chapter 3 mentioned that histograms show the shape of the distribution of a particular sample or population. In fact, if we draw histograms, we will notice that more often than not the shape looks like a bell. That means that most of the values are bunched in the center, and as you look farther and farther from the center, you find fewer and fewer observations.

In most applications and studies, distributions that have this bell shape turn out to be a particular type of bell-shaped distribution called the *normal distribution.* The normal distribution is very important in data analysis, as you will see throughout the rest of this book and in volumes 4, 5, and 6 of this series. In this chapter, you will look more closely at some characteristics of the normal distribution.

A mathematical equation defines the normal distribution exactly. For a particular mean and standard deviation, this equation determines what percentage of the observations falls where. Figure 4.1 is a picture of a normal distribution with a mean of 100 and a standard deviation of 15. As you can see, the distribution is symmetric. If you folded it in the center, the two sides would match; they are identical. The center of the distribution is at the mean. The mean of a normal distribution is also the most frequently occurring value (the mode), and it is the value that splits the distribution into two equal parts (the median). In any normal distribution, the mean, median, and mode all have the same value.

AREAS IN THE NORMAL DISTRIBUTION

For a normal distribution, the percentage of values falling within any interval can be calculated exactly. For example, in a normal distribution with a mean of 100 and

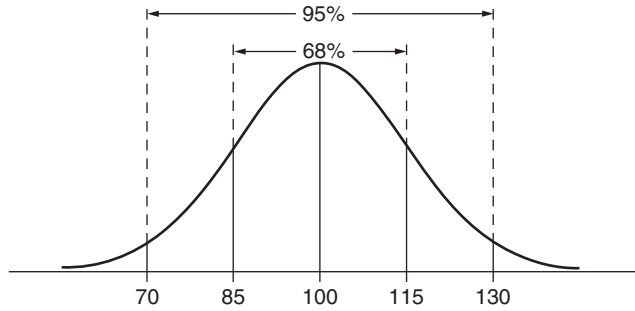


FIGURE 4.1 The normal distribution.

a standard deviation of 15 (as in Figure 4.1), 68% of all values fall between 85 (one standard deviation less than the mean) and 115 (one standard deviation more than the mean). And 95% of all values fall in the range 70 to 130, within two standard deviations from the mean.

A normal distribution can have any mean and standard deviation. However, the percentage of cases falling within a particular number of standard deviations from the mean is always the same. The shape of a normal distribution does not change. Most of the observations are near the mean, and a mathematical function describes how many observations are at any given distance (measured in standard deviations) from the mean. Means and standard deviations differ from variable to variable. But the percentage of cases within specific intervals is always the same in a true normal distribution. We are going to use this principle in Volumes IV, V and VI.

It turns out that many variables you can measure have a distribution close to the mathematical ideal of a normal distribution. We say these variables are "normally distributed," even though their distributions are not exactly normal. Usually when we say this, we mean that the histograms look like Figure 4.1. For example, Figure 4.2 shows that an actual distribution that is indeed pretty close to normal.

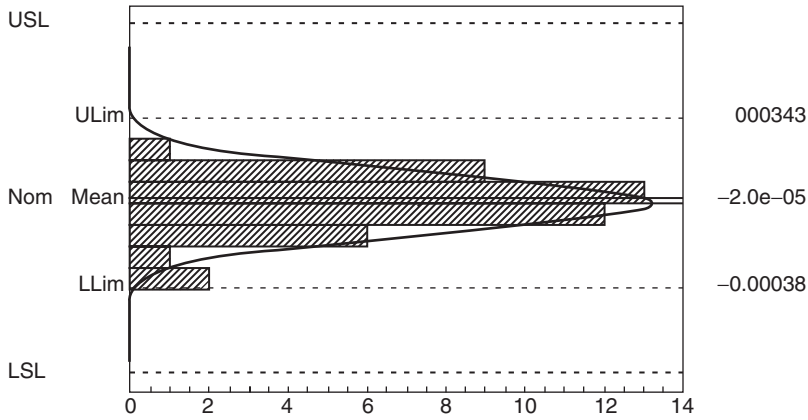


FIGURE 4.2 Comparison of actual data with a superimposed distribution curve.

STANDARD SCORES

If I tell you that I own 250 books, you probably will not be able to make very much of this information. You will not know how my library compares to that of the average consultant. Would it not be much more informative if I told you that I own the average number of books, or that I am two standard deviations above the average? Then, if you know that the number of books owned by consultants is normally distributed, you could calculate exactly what percentage of my colleagues have more books than I do.

To describe my library better in this way, you can calculate what is called a standard score. It describes the location of a particular case in a distribution: whether it is above average or below average and how much above or below. The computation is simple:

1. Take the value and subtract the mean from it. If the difference is positive, you know the case is above the mean. If it is negative, the case is below the mean.
2. Divide the difference by the standard deviation. This tells you how many standard deviation units a score is above or below the average.

For example, if book ownership among consultants is normally distributed with a mean of 150 and a standard deviation of 50, you can calculate the standard score for the 250 books I own in this way:

Step 1: 250 (my books) – 150 (average number of books) =
100 (I own 100 books more than the average consultant does.)

Step 2: 100 (difference from step 1)/50 (standard deviation of books) =
2 (standard score)

My standard score is two. Since its sign is positive, it indicates that I have more books than average. The number two indicates that I am two standard deviation units above the mean. In a normal distribution, 95% of all cases are within two standard deviations of the mean. Therefore, you know that my library is remarkable.

In a sample, the average of the standard scores for a variable is always zero, and the standard deviation is always one. Suppose you ask 15 people on the street how many hamburgers they consume in a week. If you calculate the mean and standard deviation for the number of hamburgers eaten by these 15 people and then compute a standard score for each person, you will get 15 standard scores. The average of the scores will be zero, and their standard deviation will be one.

When you use standard scores, you can compare values for a case on different variables. If you have standard scores of 2.9 for number of books, –1.2 for metabolic rate, and 0.0 for weight, then you know:

- You have many more books than average.
- You have a slower metabolism than average.
- Your weight is exactly the average.

You could not meaningfully compare the original numbers since they all have different means and standard deviations. Owning 20 cars is much more extraordinary than owning 20 shirts.

It is important to recognize that only for this example we focused on two standard deviations. Nothing prevents us from going even further. For example when we study Statistical Process Control charting we will be talking about three standard deviations, and certainly when we talk about six sigma we are indeed talking about six standard deviations.

A SAMPLE FROM THE NORMAL DISTRIBUTION

Even if a variable is normally distributed in the population, a sample from the population does not necessarily have a distribution that is exactly normal. Samples vary, so the distributions for individual samples vary as well. However, if a sample is reasonably large and it comes from a normal population, its distribution should look more or less normal. The majority of statistical packages do offer this functionality. For example, the SPSS software package provides additional information so that the experimenter can actually see the misses of the distribution. Specifically, SPSS/PC + places colons and dots on the histogram to show a true normal distribution. The colons and dots indicate how many cases would be expected in the intervals if the distribution were exactly normal with the same mean and standard deviation as the sample. A colon appears in place of the bar if the normal distribution falls inside the histogram. A dot appears if the normal distribution falls outside the histogram (beyond the end of the bar).

DISTRIBUTIONS THAT ARE NOT NORMAL

The normal distribution is often used as a reference for describing other distributions. A distribution is called skewed if it is not symmetric but instead has more cases (more of a "tail") toward one end of the distribution than the other. If the long tail is toward larger values, the distribution is called positively skewed, or skewed to the right. If the tail is toward smaller values, the distribution is negatively skewed, or skewed to the left. A variable such as income has a positively skewed distribution. That is because some incomes are very much above average and make a long tail to the right. Since incomes are rarely less than zero, the tail to the left is not so long.

If a larger proportion of cases falls into the tails of a distribution than into those of a normal distribution, the distribution has positive kurtosis. If fewer cases fall into the tails, the distribution has negative kurtosis. You can compute statistics that measure how much skewness and kurtosis a distribution has, in comparison to a normal distribution. These statistics are zero if the observed distribution is exactly normal. Positive values for kurtosis indicate that the tails of a distribution are heavier than those of a normal distribution. Negative values indicate that a distribution has lighter tails than a normal distribution does. Of course, the measures of skewness and kurtosis for samples from a normal distribution will not be exactly zero. Because of variation from sample to sample, they will fluctuate around zero. To use the

computer for the calculations, one only needs to identify the command with skewness and kurtosis, and the computer does the rest.

MORE ON THE DISTRIBUTION OF THE MEANS

It is understandable that certain variables, such as height and weight, have distributions that are approximately normal. We know that most of the world is pretty close to average and that the farther we move from average, the fewer people we find. But why does the distribution of sample means look like a normal distribution?

This remarkable fact is explained by the Central Limit Theorem. The Central Limit Theorem says that for samples of a sufficiently large size, the real distribution of means is almost always approximately normal. The original variable can have any kind of distribution. It does not have to be bell-shaped in the least. ("Real" distribution means the one you would get if you took an infinite number of random samples. The "real" distribution is a mathematical concept. You can get a pretty good idea of what the "real" distribution looks like by taking a lot of samples and examining plots of their values — as we have been doing.) Sufficiently large size? What kind of language is that for a mathematical theorem? Actually, our paraphrase of the Central Limit Theorem has several vague parts. You have to say what you are willing to consider "approximately normal" before you know what size sample is "sufficiently large." How large a sample you need depends on the way the variable is distributed. The important point is that the distribution of means gets closer and closer to normal as the sample size gets larger and larger — regardless of what the distribution of the original variable looks like. Ultimately, the means will look like a normal distribution. That is why the normal distribution is so important in data analysis. Your variable does not have to be normally distributed. Means that you calculate from samples will be normally distributed, regardless. If the variable you are studying actually does have a normal distribution, then the distribution of means will be normal for samples of any size. The further from normal the distribution of your variable is, the larger the samples have to be for the distribution of the means to be approximately normal. This is a fundamental assumption under which Statistical Process Control charting operates.

MORE ABOUT MEANS OF MEANS

So far, we have seen that for a sufficiently large sample size, the distribution of means is normal. That tells us a lot about how likely different means are, but only if we know what the mean and standard deviation of the distribution are. The mean of the "real" distribution of means is the population mean. The mean of the means is the mean? What does that mean? It means (one step at a time, now): Suppose you could take an infinite number of samples and calculate the average for each one. Suppose you could then calculate the average of your averages. What you would get is the same number as if you just went ahead and took the average of the whole population. That really is not surprising at all. For example, if 50% of all the people in a population agree with a statement, then:

- The true population mean is 50%. We just said that: 50% of all the people in a population agree.
- The mean of the distribution of sample means from that population is 50%, too.

Similarly, if the average response of a given study in the population is 100, then the mean of the distribution of means from the population is also 100. It does not matter how large the samples are, whether you have ten-case samples or 10,000-case samples. Nor does it matter whether the response is normally distributed. The mean of the distribution of means is the population mean.

THE STANDARD ERROR OF THE MEAN

If the mean of the distribution of sample means is the population mean, what is the standard deviation of the distribution of sample means? Is it also just the standard deviation of the population? No. As previously mentioned, the standard deviation of the means depends on two things:

- How large a sample you take. Larger samples mean a smaller standard deviation for the sample means.
- How much variability exists in the population. Less variability in the samples also means a smaller standard deviation for the sample means.

To calculate the exact standard deviation of the distribution of sample means, you must know:

- The standard deviation in the population.
- The number of cases in the sample.

All you have to do is divide the standard deviation by the square root of the sample size. The result, the standard deviation of the distribution of sample means, is called the *standard error of the mean.* Although it has an impressive name, it is still just a standard deviation — the standard deviation of the sample means. Think about the formula for computing the standard error of the mean: take the standard deviation of the variable and divide by the square root of the sample size. Suppose the standard deviation of number of books owned is 50, and the sample size is four cases. Then the standard error is 50 divided by the square root of 4, to yield 25. If the sample size is increased to 9, the standard error decreases to 50 divided by the square root of 9, or 16.7. If the sample size is increased to 100, the standard error is only 5. The larger the sample size, the less variability there is in the sample means.

CALCULATING A CONFIDENCE INTERVAL

You have spent a lot of time reading about sample means and how they vary. At this stage you may wonder why this is necessary. The reason is very simple. You have

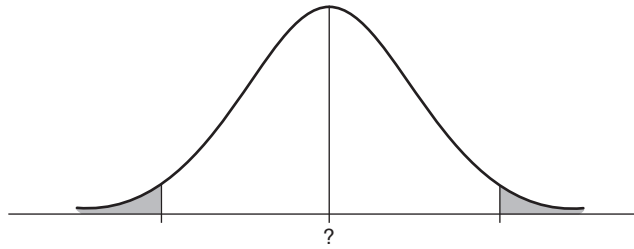


FIGURE 4.3 The sampling distribution of means.

to understand these things in order to use statistics for testing hypotheses about the population. If you know how much the means vary from sample to sample, you can draw conclusions about the population by looking at just a single sample. Watch.

Take a well-defined population — the owners of a particular product X. Suppose that you want to estimate the average satisfaction with that purchase. You randomly select 25 individuals who have returned their registration cards and send them a questionnaire. The average satisfaction response of these 25 customers turns out to be 112, and the standard deviation of their response scores is close to 15, the value for the population. Based on this sample, what can you conclude about all of the customers who have purchased product X?

The sample you selected is one of many possible samples. So the mean you calculated is one of many possible means. In particular, it is one of the means in the distribution of means for samples of size 25. The problem is that you do not know where your sample falls in the distribution of means. Is it close to the true population value? Is it one of the extreme means? Since you do not know the true value for the response of people who have purchased the product, you cannot tell if your sample value is too high, too low, or right on target. You never know the true value in the population, because if you did you would not do the study.

You do not know the population mean, and therefore you do not know the mean of the distribution of sample means. Nevertheless, you can estimate the standard error of the mean from your observed standard deviation. Remember, the standard error of the mean is the standard deviation of the distribution of sample means. The estimated standard error is the standard deviation (15) divided by the square root of the sample size (25), which makes 15 divided by 5, or 3. Using this piece of information, you can visualize the sampling distribution of means, as shown in Figure 4.3.

Based on the Central Limit Theorem, you can assume that the distribution is normal. That is what the Central Limit Theorem says: for a sufficiently large sample size, sample means are normally distributed whether the original variable (satisfaction in this case) is normally distributed or not. Since you do not know the mean satisfaction for the population of owners of the product X, it is labeled with a question mark in the figure. Because the distribution is normal, you know that 95% of all sample means should fall within two standard errors of the mean. The standard error of the mean was found to equal three. So 95% of all sample means should fall within six of the question mark. The values falling outside of this interval are shaded in the figure. (We can do this for any level of standard error.)

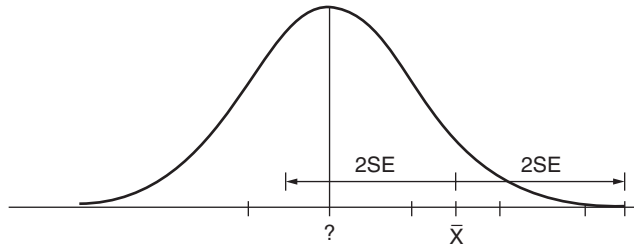


FIGURE 4.4 The sample mean 1.5 standard errors above the population mean.

Where is your sample mean in this distribution? Sorry — you cannot figure that out. If you knew the population value (at the question mark), then you could mark the location of the mean; but, of course, you do not. Based on this picture, what can you say about the value of the population mean? Although you cannot give an exact value, you can calculate a range of values — an interval that should include the population mean 95% of the time. You calculate the lower limit of this interval by subtracting two times the standard error from your mean. The lower limit is therefore 112 – 6 = 106. You calculate the upper limit by adding two times the standard error to your mean. This is 112 + 6 = 118. The interval is from 106 to 118. Now you have what is known as a *confidence interval,* extending from two standard errors below the sample mean to two standard errors above the sample mean.

Think of what the diagram shows. You can imagine your sample mean somewhere in the distribution and see what happens. Figure 4.4 shows the sample mean at 1.5 standard errors above the population mean (the question mark). The confidence interval is marked off. Does the interval include the unknown population value? Sure — because it reaches out two standard errors, and the difference between your sample mean and the population mean is only 1.5 standard errors.

Now imagine your sample value at one standard error unit below the mean, as in Figure 4.5. Does the confidence interval still include the population value? Yes. Once again, your sample mean is within two standard errors of the population mean, so the population mean lies within the confidence interval. The only time your interval would not include the population value is when your sample mean falls in the shaded area of Figure 4.3. The shaded region corresponds to the 5% of the distribution that is more than two standard error units from the population mean.

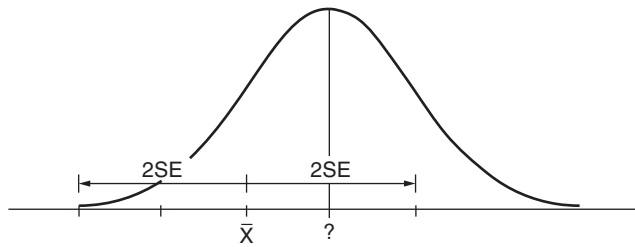


FIGURE 4.5 The sample mean 1 standard error below the population mean.

(This is the region most often used for analysis. However, in some cases this is not good enough and we have to go out to three even to six standard error units.)

You do not know the exact value for the population mean. But as shown here, you can calculate an interval around your sample mean that will include the true, unknown population mean 95% of the time. This is called a 95% confidence interval. Of course, you can never tell whether your particular sample mean is one of the unlikely ones in the shaded region. All you can do is calculate the interval and hope that you have one of the 95-out-of-a-100 times that the interval includes the population value. By the way, there is nothing sacred about the 95% confidence. A given confidence is determined a priori by the experimenter and depends on the practicality of the study. So it is not unusual to see a 90%, 99%, 99.9% or any other confidence.

MORE SATISFIED THAN AVERAGE?

You calculated a mean satisfaction response of 112 for a sample of 25 customers for product X. This is 12 points higher than 100, which is supposedly the average value for people in general. Is it reasonable to conclude that these customers of product X are different, on average, from people in general? You know that sample means vary, so you do not expect the value observed in a sample to be exactly the same as the population value. And now the question arises: Where did the population come in here? We were just looking at people who purchased product X, right? Yes. In this rather improbable study, the "population" is just the people selected for the study based on the most current information about the purchasers of this particular product. For them, the "population value" of average satisfaction response is the value you would get if you gave a questionnaire to everybody who bought the product and you then calculated the average. The problem with this scenario, however, is that any time you are talking about statistics, the word "population" has a special meaning. It is the people (or animals or things) that you are trying to draw conclusions about. In this study, it is the satisfaction level of all possible customers for this product. In other words, we use samples to predict populations in our simple case. What you have to figure out from your sample of respondents is this: How likely is this sample mean of 112 if the population mean (for everybody who ever bought this product X) is 100?

Use a picture again. Figure 4.6 is the distribution of means for samples of size 25 when the population value is 100 and the standard deviation is 15. It looks a lot

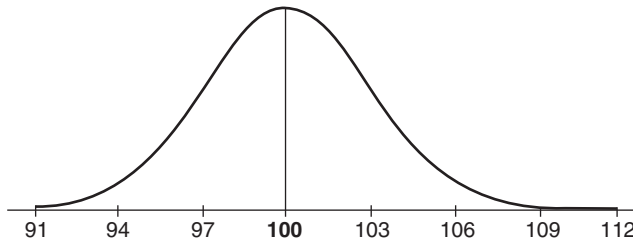


FIGURE 4.6 The distribution of mean for sample size 25.

like some of the previous diagrams. The difference is that instead of the question mark, you see the value 100, the mean for people in general. You can now locate the observed sample mean on the distribution. It does not fit well, since the value 112 is four standard error units above the mean. This result indicates that it is very unlikely to observe a sample mean as large as 112 in a sample of size 25 when the true population value is 100. Only about 0.006% of the cases in a normal distribution have values as much as four standard deviations away from the mean. So it appears highly unlikely that the customers of product X have the same mean satisfaction response as the general population. On the other hand, if you had observed a mean response of 103 in your sample, you could not say with confidence that the customers were more satisfied than average, since a sample mean of 103 is perfectly reasonable for a population value of 100.

5 Testing Hypotheses about Two Independent Means

The previous chapters have addressed several issues about samples, means, and standard errors, with the intent to explain the data. The implication was that they are all useful but quite different from each other. In fact, you may get different answers if you use different samples. This chapter will explore these differences because the real question is, How much will they differ? How can you decide whether a difference in sample means can be attributed to their natural variability or to a real difference between groups in the population?

OVERVIEW

At this point in the book, you can look at specific data and describe a sample. That is all you can do if you cannot understand the relationships between samples and populations. However, now you are ready to do more. You can look in the sample at the percentage of a response and relate that information back to the population. You know that you have one of many possible samples and that the chances are slim that the value calculated from the sample is identical to the population value. You can also calculate the standard error and use that number to calculate a 95% confidence or any other confidence you desire. Even though this calculation is possible, you never know if the particular interval you calculated contains the population value (it either does or it does not). However, you do know that the interval will include the population value 95 times out of 100 (or whatever the confidence).

If the confidence is narrow, that is a good indication because you want to pinpoint the population value as closely as possible. You do not know where within a confidence interval the population value might be. It is much more useful, then, to know that the 95% confidence interval is between a set of limits rather than outside of either side of the set. When you conduct a survey or experiment, look at both the mean and its confidence interval. If the interval is wide, you have only a very rough estimate of the population mean.

IS THE DIFFERENCE REAL?

So now the question is, "Is the difference real?" Usually when you conduct a study, you have some ideas that you want to explore. These ideas, often called hypotheses, typically involve comparisons of several groups, such as "Do men and women find life equally exciting?" "Does income differ between people who find life exciting and those who do not?" "Is a new machine making a real difference?"

Chances are if you compare two or more things, you are going to find some differences. After all, no two things are exactly the same. The question is not so much if things are different but rather, what can you make of the difference?

So far we have seen that different samples from the same population give different results. The real issue is, how much will they differ? How can you decide whether a difference in sample means can be attributed to their natural variability or to a real difference between groups in the population?

EVALUATING A DIFFERENCE BETWEEN MEANS

How can you decide when a difference between two means is big enough for you to believe that the two samples are from a population with different means? It depends on how willing you are to be wrong. Look at Figure 5.1, which is the real distribution of differences for samples of size 20 from a distribution with a standard deviation of 15. (You can calculate the standard deviation of the distribution of differences. It is called the standard error of the difference.) Since the distribution is normal, you can find out what percentage of the samples falls into each of the intervals.

The scale on the distribution is marked with the actual values and with "standardized" values, which are computed by dividing the differences by the standard error. Looking at standardized distances is convenient, since the percentage of cases within a standardized distance from the mean is always the same. For example, 34% of all samples are between zero and one standardized unit greater than the mean, and another 34% are between zero and one standardized unit less than the mean. If you always express your distances in standardized units, you can use the same normal distribution for evaluating the likelihood of a particular difference.

From Figure 5.1, you can see that about 13% of the time, you would expect to have at least a 7-point difference in the sample means when two population means

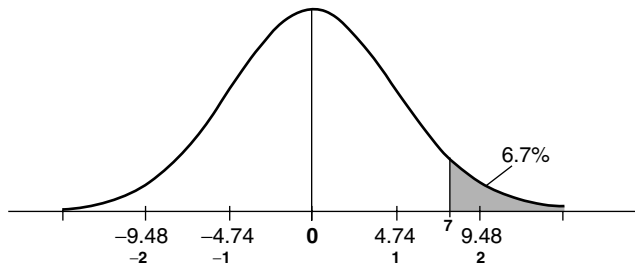


FIGURE 5.1 Theoretical distribution of differences of means.

are equal. Why? You just found out that the 7-point difference is 1.5 standard errors. Look on the figure to see what percentage of the differences is that big. You should look at the area to the right of +1.5 and the area to the left of –1.5. Since the distribution is symmetric, the two areas are equal. Each one is about 6.7% of the total, and together they make up a little over 13% of the total. So about 13% of differences in means are going to be as big as 1.5 standard errors (or 7 points) if the real difference in means is zero.

WHY THE ENTIRE AREA?

You may wonder why you do not find the probability of getting a difference of just seven. Think of the following analogy. You are tired of the life of a poor student and have decided that the quickest (legal) way to upgrade your status is to marry rich. You settle on a definition of rich. Perhaps you need an income of $250,000 a year. Now you want to see how likely it is that you can achieve your goal. You go to the university library and ask the reference librarian to find some facts. Would you ask about just the number of eligible singles of the opposite sex with incomes of $250,000? No, you would ask for incomes of $250,000 or more, since they all satisfy your criterion of richness. In evaluating your chances of marrying rich, you would include all incomes of $250,000 or more. Similarly, when you are trying to decide whether seven is a likely outcome for a difference, your interest is not just in the number seven but in all differences that are at least that large.

Since both outcomes were possible and you did not know the outcome before you actually ran your experiment, when you evaluate the chances of seeing a difference at least as large as seven points, you have to look in both directions. Both of the extreme regions of the distribution are atypical. Sometimes, though, you can look in just one direction. It really depends on how you stated your initial hypothesis. If you hypothesized that you expect higher values you would be looking at the right tail, if lower you would be looking at the left tail of the distribution. This type of test is called a *one-tailed test.* Your decision to use a one-tailed test is a very important one because once you see the results you cannot go back and switch sides and apply the one-tailed test for a difference that is in the other direction. Use a one-tailed test only if you definitely expect one specific group to be higher. Otherwise, use a *two-tailed test* and look at both sides of the distribution.

DRAWING A CONCLUSION

Now that you have seen how to use the standard error of a difference in means, how do you compute it? You take the square root of the variance of the difference. How do you get the variance of the difference? When you have two means from independent samples, the variance of their difference equals the sum of their variances. This neat little fact would take too long to prove here, but you can see how it is used. The example above had two means from independent samples of size 20, taken from a population whose standard deviation was 15. You calculate the standard error of the difference by following these steps:

- As explained earlier, the standard error of each mean equals 15 (the standard deviation) divided by the square root of 20 (the sample size).
- The variance of each mean is just the square of that fraction: 15 squared divided by 20. That is 225 divided by 20, or 11.25.
- The variance of the difference between the means is the sum of the variances of each mean. That is 11.25 plus 11.25, or 22.5.
- The standard error of the difference between the means is therefore the square root of 22.5, or 4.74.

MORE ON HYPOTHESIS TESTING

In the previous example, you performed a statistical test of a hypothesis. You tested whether two variables have the same average values in the population. This was the basic procedure:

- You wanted to draw conclusions about the population, but you could not experiment with the entire population, so you had to base your conclusions about the population on the results from the sample.
- You calculated how likely it is that a difference as large as the one you observed would occur if no difference exists between the two means in the population.
- Since there was a 13% chance that you could see a difference as large as the one you observed if the population means did not differ, the evidence was too skimpy to reject the hypothesis of no difference.

WHY IS THAT SO COMPLICATED?

You are probably wondering why all of a sudden we use verbiage that perhaps is not very scientific. Why are we going around in circles? Why do we have to assume that no difference exists between the means in the population and then figure out how likely the observed results are if no difference exists? Why not just calculate the probability that a difference exists? That is what we really want to know, is it not?

Although it sounds like a good idea, in this situation you cannot calculate the probability that a difference is present. A difference either exists or it does not. If we have two sample means, say 11 and 12, what do they tell us about whether it is true that the two means in the population are different? Not much. The probability of getting two sample means that differ by at least one depends on how much of a difference is present in the population. It depends on whether the true difference is 0, 1, 2, 4, 100, or whatever. A difference of one may be very unlikely if the true difference is 100 but perfectly likely if the true difference is zero or two. But we do not know what the true difference is. We can only consider the likelihood of a value of one or more in relation to some hypothetical situation, such as a true difference of zero or a true difference of 100. We cannot assign the difference some overall probability.

What if you found that the two sample means were 11 and 11? Would you claim that it is certain the two means are exactly equal in the population? Would you be

willing to forget the possibility that the population means might be 11 and 11.1? Of course not. (I hope). You have seen that samples vary and that it is most unlikely for two sample means to be exactly equal even if the two means are equal in the population. Similarly, you can easily get sample means that are the same from populations whose means differ to a small extent.

You simply cannot figure out the probability that two population means are equal or unequal. You can, however, estimate the probability that you would see a difference of at least two (or some other value) in the sample when no difference exists in the population (or when a difference of a particular size is present). In the previous example, you saw the calculations for the probability that the means from two samples would differ by at least seven when no difference exists in the population.

To test a hypothesis, you do the following:

1. State the hypothesis of interest. This is what you think is really true for the population.
2. Determine the frame of reference you will use to evaluate your hypothesis. This is what is true in the population if your hypothesis is wrong. This "frame of reference" is called the *null hypothesis,* since it describes the population when the hypothesis you are interested in is not true, when it is null.
3. Calculate the probability that you would see a difference at least as large as the one you observed in your sample if the null hypothesis is true.
4. If this probability (called the observed significance level) is small, say less than .05, reject the null hypothesis.
5. If the observed significance level is large, do not reject the null hypothesis. This does not mean that you accept the null hypothesis. You simply do not reject it. You remain uncertain.

You must state the null hypothesis in a way that allows you to calculate the distribution of sample means when it is true. You cannot use a null hypothesis that says the population means are unequal, since no single distribution of sample means exists for that statement. But you can have a null hypothesis that says the difference between two population means is five or some other particular number. The null hypothesis must provide the reference point for calculating the probability of the observed results. You calculate the probability of the observed results if the null hypothesis is true.

6 Testing Hypotheses about Two Dependent Means

The previous chapters have addressed several issues about samples, means, and standard errors, with the intent to explain the data. It has been pointed out that all differ and that depending on the sample you may get different answers. This chapter will explore these differences based on two dependent means because the real issue is, how much do they differ? How can you decide whether a difference in sample means can be attributed to their natural variability or to a real difference between groups in the population? And what is significant?

OVERVIEW

Consider another example of testing a hypothesis. We want to see if a difference exists in the average of two groups. When we do a comparison of two groups, we must also be cognizant of the intervening variable(s). An intervening variable is a variable that may explain the difference between the two groups. If that is the case we can minimize this problem by restricting the comparison of that particular variable for that group. By doing that restriction we will no longer affect the comparison as much. As a consequence, the results will be much easier to interpret.

In this case, the null hypothesis is that the groups have the same average. On the other hand, the hypothesis of interest (sometimes called the *alternative hypothesis*) is that the average of the two groups is not the same. The groups differ.

USING THE t DISTRIBUTION

So far, we always knew or pretended to know the standard deviation in the population. In fact, though, it usually must be estimated from the sample. When this is necessary — when we use the same sample both to test the hypothesis and to estimate the standard deviation in the population — we have to use the t distribution instead of the normal distribution. The t distribution is much like the normal distribution. It just shifts the area in the normal distribution to adjust for the fact that we do not know what the standard deviations really are. (When sample sizes are large, the t distribution looks very much like the normal distribution.)

As always, we compute the difference between the two means, find its standard error, and then calculate how improbable the observed difference is. However, the answer to our question of "significant" difference is dependent on the pooled variance estimate and the degrees of freedom. The degrees of freedom are based on the number of observations in each of the two groups.

TWO TYPES OF ERRORS

You can make two types of mistakes when testing a hypothesis about two means. You can claim that the two means are not equal in the population when in fact they are. Or you can fail to say that a difference exists when it actually does. Statisticians, being very methodical people, have given these two types of errors particularly descriptive, easy-to-remember names. They call the first error (claiming that two means are not equal when in fact they are) a *Type 1 error.* The second type of error (not finding a difference when one really exists) is called a *Type 2 error.*

It may be easy to remember that you call the two kinds of error Type 1 and Type 2, but how do you remember which is which? Perhaps you can remember it this way. The Type 1 error is the error you are tempted to make. When you say, proudly, "There is a difference. Something is happening here. I have found a relationship," you are taking the chance of making a Type 1 error.

If you can remember what the Type 1 error is, then it is pretty easy to figure out that the Type 2 error is the one you are not tempted to make, saying "Nothing is happening here" when a difference really is present in the population.

Using Software to Generate the Output from the t Test Procedure

The computation of the t test differs depending on whether you assume that in the population the two groups have the same variances or not. If you can assume that the two variances are equal, use the numbers in the columns labeled pooled variance estimate. If you cannot assume that the two variances are equal, use the t test labeled separate variance estimate. The ratio of the variances in the two samples is shown in the column labeled F value. Next to the F value, most software shows the probability that you would see a difference at least as large as the one observed in the sample if the variances are equal in the population and if the distribution of the variable is normal. (The F test for equality of variances is quite sensitive to departures from normality, while the t test is not. If the data are not from normal populations, the observed significance level for the F statistic may be unreliable.) If the observed significance level is large, you have little reason to worry about your variances. If the number is small, you should use the t test marked separate variance estimate. In general, it is a good idea to use the separate variance t test whenever you suspect that the variances are unequal.

INTERPRETING A t TEST

It is impossible to prove, based on samples, that two population means are exactly equal. What if variable X in the population has a mean of 3.2500 units, and variable

Y has a mean of 3.2501 units? Since sample means differ, and the statistical procedures for evaluating differences between means must allow for variability from sample to sample, we will never be able to detect such a small difference in the population. What does happen instead is that we take two samples, compute a t test, and find a large observed significance level. Perhaps we find a probability of .50 that the t value could be observed in a population with no difference. This large observed significance level does not tell us that the means are exactly equal. It just indicates that the results would not be "far out" if the two means are equal in the population. So instead of embracing the null hypothesis and claiming that it is true, we just say that we have no evidence to believe that it is not true. We cannot prove the null hypothesis.

AN ANALOGY: COIN FLIPS

Suppose someone comes up to you, hands you a coin, and says, "Tell me if this is a fair coin — a coin for which heads and tails are equally likely." If you had nothing better to do, you would probably start flipping the coin and counting the number of times heads and tails occur. But you are no longer naive. You know that if you flip a fair coin 10 times, you will not often get exactly five heads and five tails. All sorts of outcomes are possible. There is even a reasonable chance that with a fair coin you will get eight tails and two heads or eight heads and two tails. However, as the coin becomes less and less unfair, it gets harder and harder for you to detect the difference. If the true probability of a head on the coin is .4999 instead of .5000, you would never figure that out unless you are willing to spend the rest of your life flipping coins. Any combination of flips that you come up with will appear perfectly reasonable if the coin is fair or if it is minutely biased. Although you can disprove with a certain degree of confidence that a coin is fair, it is impossible to prove that it is exactly fair. That is why we cannot say we proved that the average of any variable or group is the same as the average of another variable or group. All we can say is that the evidence did not disprove it.

OBSERVED SIGNIFICANCE LEVELS

When your observed significance level is small, its interpretation is fairly straightforward: the two means seem to be unequal in the population. The observed significance level tells you the probability that the observed difference could be due to chance. The observed significance level is the probability that your sample could show a difference at least as large as the one that you observed if the means are really equal.

So what is a small significance level? Most of the time, significance levels are considered small if they are less than .05; sometimes, if they are less than .01. Rather than just rejecting or not rejecting the null hypothesis, look at the actual significance level as well. An observed significance level of .06 is not the same as an observed significance level of .92, though both may not be statistically significant. When reporting your results, give the exact observed significance level. It will help the

reader evaluate your results. Treat the observed significance level as a guide to whether or not the difference could be due to chance alone.

If your observed significance level is too large to reject the hypothesis that the means are equal, more than one explanation is possible. The first explanation is that no difference may exist between the two means or that it may be so small that you cannot detect it. If the true difference is very small, it may not matter that you cannot find it. Who really cares about a tiny difference (such as a difference in annual income of ten dollars)? Little, if anything, is lost by your failure to establish such tiny differences.

The second explanation is more troublesome. Perhaps an important difference does exist, and you cannot find it. This can occur if the sample size is small.

If you flip a coin only twice, you cannot establish whether it is fair. A fair coin has a 50% chance of coming up heads twice or tails twice in two flips and a 50% chance of coming up with one of each. Any outcome that you see is consistent with the coin's being fair. As the number of flips increases, so does your ability to detect differences. To detect a small difference, you need a big sample so that the difference would clearly be outside the expected degree of sample variation.

The variability of the responses (in the population) also affects your ability to detect differences. If the observations vary a great deal, the sample means will vary a lot as well. Even large differences in observed means can be attributed to variability among the samples.

To wrap up all of this: if you do not find evidence to reject the hypothesis that two means are equal in the population, one of two possibilities is true:

- The means are equal or very similar.
- The means are unequal, but you cannot detect the difference because of small sample size, large variability, or both.

TAILS AND SIGNIFICANCE TESTS

An observed significance level printed on a computerized t test output is labeled 2-TAIL PROB, standing for two-tailed probability. This value tells you the probability that you would see in either direction a difference at least as large as the one you would observe if no difference existed in the population. Either the first group has a mean larger than that of the second group by at least the observed size, or the second group has a mean larger than that of the first group by at least the observed size.

If you do not know which of the two groups should have the larger mean, that is what you have to ask. Differences in either direction cast doubt on the null hypothesis that in the population the two groups have the same means. If you do know in advance which group will have the larger mean if they differ, then you use a one-tailed significance level.

Suppose you know that a new drug for insomnia will either leave the length of time you need to fall asleep unchanged or decrease it. You take two random samples of people and perform an experiment. One group gets the drug, and the other gets a placebo (a fake drug just to make the subjects think they are being treated). Then you find the average time it takes each group to fall asleep. You calculate the

difference between the two means, along with its standard error. To find out how often you would get a difference of this magnitude by chance when the drug and placebo are equally effective, you need only calculate the probability that you see a decrease at least as large as the one observed. You are confident that people treated with the drug will not take longer to fall asleep, so you decide in advance not even to test for that possibility.

Think back to the coin analogy. Suppose your friend tells you, as he is handing you the coin, that he suspects the coin is biased. Perhaps his wife always uses it to settle disputes, and she always bets heads and wins. You have a pretty good suspicion that the coin is biased in favor of heads and no reason at all to suspect bias in favor of tails. If you toss the coin ten times, and it comes up all heads, you just want to know what the probability is that a fair coin comes up ten heads out of ten flips. You do not worry that it might have come up with ten tails, since the only situations that will cause you to doubt your coin are excesses of heads.

If you know in advance which of two means should be larger, you can convert the two-tailed significance level to a one-tailed level. All you do is divide the two-tailed probability by two. The result tells you the percentage of the t distribution in one of the tails.

THE HYPOTHESIS-TESTING PROCESS

In the previous example, we used a statistical technique called the t test to test the hypothesis that two groups have the same mean in the population. We did the following:

- For each of the groups, we calculated the mean of the variable we were interested in comparing.
- We subtracted one mean from the other to determine the difference between the two.
- We calculated a t statistic by dividing the difference of the two sample means by its standard error.
- We calculated the observed significance level. This told us how often we would expect to see a difference as large as the one we observed if no difference existed between the groups in the population.
- If the observed significance level was small (less than .05), we rejected the hypothesis that the two means are equal in the population.
- Otherwise, we did not reject the null hypothesis, and we did not accept it either. We remained undecided. That is because we did not know whether no difference was present or whether our sample was simply too small to detect the difference.

This procedure is the same for tests of most hypotheses:

- You formulate a null hypothesis and its alternative.
- You calculate the probability of observing a difference of a particular magnitude in the sample when the null hypothesis is true.

- If this probability (the observed significance level) is small enough, you reject the null hypothesis.
- If the probability is not small enough, you remain undecided.

The only part of this ritual that changes for different situations is the actual statistic used to evaluate the probability of the observed difference. In the chapters that follow, we will use different types of statistics to test hypotheses that have these characteristics:

- Variables are independent.
- Several groups have the same means.
- No linear relationship exists among several variables.

If you make sure now that you understand the way hypothesis testing works, the rest of this book will be easy to understand.

ASSUMPTIONS NEEDED

To perform a statistical test of a hypothesis, you must make certain assumptions about the data. The particular assumptions you must make depend on the statistical test you are using. Some procedures require stricter assumptions than others. The assumptions are needed so that you (or your computer) can figure out what the distribution of the statistic is. Unless you know the distribution, you cannot determine the correct significance levels. For the pooled-variance t test, you need to assume that you have two random samples with the same population variance. You also need to assume that the distribution of the means is approximately normal, which can happen one of two ways:

- The variable is normally distributed, so the means will automatically be normally distributed.
- The sample size is large enough to allow you to rely on the Central Limit Theorem to make sure that the means are distributed normally.

Of course, some assumptions are more important than others. Moderate violation of some of them may not have very serious consequences. Therefore it is important to know, for each statistical procedure, not only what assumptions are needed but also how severely their violation may influence the results. We will talk about these things when we discuss the different statistical procedures. For example, as mentioned earlier, the F test for equality of variances is quite sensitive to departures from normality. The t test for equality of means is less so.

Based on the means observed in two independent samples, how can you test the hypothesis that two population means are equal? Here is the procedure:

- To test the null hypothesis that two population means are equal, you must calculate the probability of seeing a difference at least as large as the one you have observed if no difference exists in the population.

- The hypothesis that no difference exists between the two population means is called the null hypothesis.
- The probability of seeing a difference at least as large as the one you have observed, when the null hypothesis is true, is called the observed significance level.
- If the observed significance level is small, usually less than .05, you reject the null hypothesis.
- If you reject the null hypothesis when it is true, you make a Type 1 error. If you do not reject the null hypothesis when it is false, you make a Type 2 error.
- The t test is used to test the hypothesis that two population means are equal.

PAIRED EXPERIMENTAL DESIGNS

When comparing two treatments in an experiment, is it always better to form pairs of similar subjects (or to observe the same subject under both conditions) than to use two independent groups? No. Paired experimental designs are only useful when you can form pairs on the basis of a variable that is related to the one you are studying. If you pair your subjects based on shoe size when you are studying responses to a new drug, the paired design actually makes it less likely that you can identify a true difference when it exists. The two members of a pair are not alike in any way that matters. In this case, using a paired t test makes it more difficult, statistically, to detect true differences than just treating the two groups as independent samples.

A paired design is a good way to eliminate some of the differences between subjects in two groups so you can focus on the particular difference that you are testing. But you need to keep some things in mind:

- If the effect of a treatment does not wear off quickly, you must make sure that enough time passes between treatments so that one wears off before another begins. Otherwise, you will not know whether the first or the second treatment is causing the results during the second observation.
- You should also be aware of the learning effect. You encounter it when a subject's response improves merely by doing the same thing again. For example, if you give subjects the same test twice, they may do better the second time, regardless of what else has happened.

You must pay attention to both the timing and the sequence of administering the treatments. You may also want to include a control group that receives no treatment but undergoes the same measurements that the treatment groups undergo.

SIGNIFICANCE VS. IMPORTANCE

What does it mean if you reject the null hypothesis that two population means are equal? Does it mean that an important difference exists between the two groups? Not necessarily. Whether a difference of half a year of education between two groups

is found to be statistically significant or not depends on several factors. It depends on the variability in the two groups, and it depends on the sample sizes.

A difference can be statistically significant with a sample size of 100, while the same difference would not be significant with a sample size of 50. The difference between the two sample means is the same, half a year, but its statistical interpretation differs. For large sample sizes, small differences between groups may be statistically significant, while for small sample sizes, even large differences may not be.

What can you make of this? Finding that a difference is statistically significant does not mean that the difference is large, nor does it mean that the difference is important from a research point of view. For sufficiently large sample sizes, you might find that even a small difference is statistically significant. That does not mean that the difference is of any practical importance. For example, we all know that on the average, an extra month of education will not do much for you. It is unlikely to alter your perception of the world. It probably does little to enhance your ability to explain why some people find life exciting and others do not.

On the other hand, if the sample sizes in the two groups are small, even a difference of four years of education may not appear to be statistically significant. In this case, you do not want to rule out a variable that may prove to be an important variable. Instead, you must worry about the fact that with small sample sizes you can miss important differences. You must allow yourself the possibility that a big difference exists because your ability to find it is poor. The probability of detecting a difference of a particular magnitude when it exists is called the power of a test. You can estimate in advance how big a sample you need in order to detect a difference that you consider really important. Discussion of how that is done is a little beyond this book. However, the reader should understand that the power is actually determined by three factors:

Effect size — The probability of achieving statistical significance is based not only on statistical considerations but also on the actual magnitude of the effect of interest.

Alpha (α) — As alpha (Type 1 error) becomes more restrictive, power decreases. This means that as the experimenter reduces the chance of finding an incorrect significant effect, the probability of correctly finding an effect also decreases.

Sample size — At any given alpha level, increased sample sizes always produce greater power of the statistical test. But increasing sample size can also produce "too much" power. This means that when you increase sample size, smaller and smaller effects will be found to be statistically significant, until at very large sample sizes almost any effect is significant. Remember that small sample sizes will make the statistical test insensitive and that large samples will make it overly sensitive. Figure 6.1 shows an example of the impact of sample size on power for various alpha levels (.01, .05, .10). Needless to say, the experimenter must consider the impact of power before selecting the alpha level.

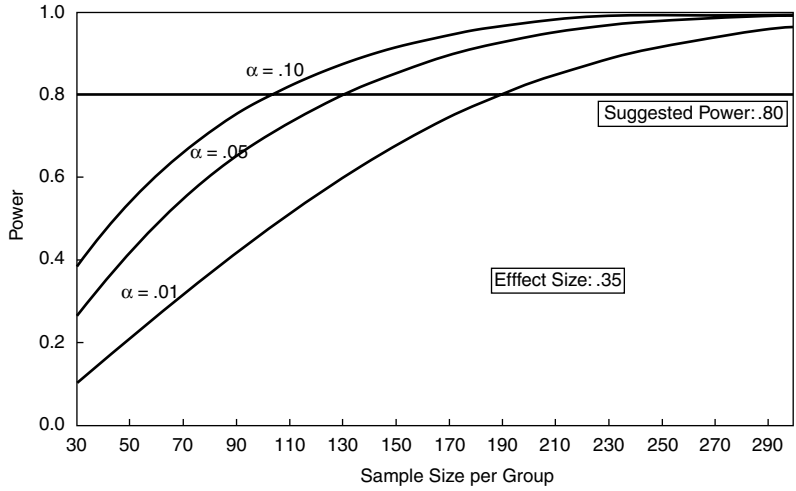


FIGURE 6.1 Impact of sample size on power for various alpha levels (.01, .05, .10).

In summary, remember that even though two groups are found to be statistically different, their difference is not necessarily of practical importance. Evaluate the difference on its own merits.

How can you test the null hypothesis that two percentages are equal in the population? How can you test the null hypothesis that two variables are independent? Here are the things that you need to know and will be discussed later:

- Observed frequencies are simply the numbers of cases with specific combinations of values.
- Expected frequencies are the numbers of cases that would have specific combinations of values if the null hypothesis were true.
- The chi-square statistic is based on a comparison of observed frequencies with expected frequencies. From it, you can obtain an observed significance level for the hypothesis that two proportions are equal.
- Two variables are independent if knowing the value of one variable tells you nothing about the value of the other.
- The degrees of freedom of a cross-tabulation reflect the number of cells in the table that are free to vary. You compute them by taking the number of rows minus one and multiplying that by the number of columns minus one.
- From the chi-square statistic and the degrees of freedom in a cross-tabulation, you can calculate the observed significance level for the null hypothesis that the two variables are unrelated.
- Chi-square increases in direct proportion to sample size if the strength of the relationship stays the same. If you double the number of cases in each cell of a cross-tabulation, chi-square is doubled.

7 Comparing Several Means

This chapter will examine whether or not the observed differences in the samples may be attributed to just the natural variability among sample means or whether a reason exists why the groups have different means in the population.

OVERVIEW

If by constructing histograms and calculating some basic descriptive statistics we have concluded that some differences are present in our groups of study, we now need to figure out whether the observed differences between the samples may be attributed to just the natural variability among sample means or whether we have reason to believe that the groups have different means in the population.

The null hypothesis says that in the population, the means of the groups are equal. That is, no difference exists in the average of our tested response. The alternative hypothesis is that a difference exists. The alternative hypothesis does not say which groups differ from one another. It just says that the groups are not all the same — at least one of the groups differs from the others.

ANALYSIS OF VARIANCE

The statistical technique used to test the null hypothesis that several population means are equal is analysis of variance (ANOVA). It is called that because it examines the variability in the sample and, based on the variability, it determines whether there is reason to believe the population means are unequal. We will be drawing conclusions about means by looking at variability.

All software statistical packages contain several different procedures that can perform analysis of variance. We will begin with the simplest format of ANOVA, which is the ONEWAY procedure. It is called one-way analysis of variance because cases fall into different groups based on their values for one variable.

NECESSARY ASSUMPTIONS

The data must meet two conditions for you to use analysis of variance:

1. Each of the groups must be a random sample from a normal population.
2. In the population, the variances in all groups must be equal.

You can visually check these conditions by making a histogram of the data for each group and seeing whether the data are approximately normal. To check whether the groups have the same variance in the population, you can examine the histograms as well as compute the variances for each of the groups and compare them. In practice, analysis of variance gives good results even if the normality assumption does not quite hold. If the number of observations in each of the groups is fairly similar, the equal variance assumption is also not too important. The assumption of random samples, however, is always important and cannot be relaxed.

In analysis of variance, the observed variability in the sample is divided, or partitioned, into two parts: variability of the observations within a group (around the group mean) and variability between the group means. Why are we talking about variability? Are we not testing hypotheses about means? Previously we mentioned that a relationship exists between variability of observations (in the population) and variability of sample means. If you know the standard deviation of the observations, you can estimate how much the sample means should vary. In your study, you have several different groups (for example, in our earlier example of the ride, we had individuals who found it comfortable, normal and rough). If the null hypothesis is true (that is, if all three groups have the same mean in the population), you can estimate how much observed means should vary due to sampling variation alone. If the means you actually observe vary more than you would expect from sampling variation, you have reason to believe that this extra variability is due to the fact that some of the groups do not have the same mean in the population.

WITHIN-GROUPS VARIABILITY

We are going to look a little more closely now at the two types of variability we need to consider. *Within-groups variability* is a measure of how much the observations within a group vary. It is simply the variance of the observations within a group in your sample, and it is used to estimate the variance within a group in the population. (Remember, analysis of variance requires the assumption that all of the groups have the same variance in the population.) Since you do not know if all of the groups have the same mean, you cannot just calculate the variance for all of the cases together. You must calculate the variance for each of the groups individually and then combine these into an "average" variance.

For example, suppose you have three groups of 20 cases each. All 20 cases in the first group have a value of 100, all 20 cases in the second group have a value of 50, and all 20 cases in the third group have a value of 0. Your best guess for the population variance within a group is zero. It appears from your sample that the values of the cases in any particular group do not vary at all. But if you had computed the variance for all of the cases together, it would not even be close to zero. You would calculate the overall mean as 50, and cases in the first and third groups would all vary from this overall mean by 50. There would be plenty of variation.

BETWEEN-GROUPS VARIABILITY

We noted earlier that a relationship exists between the variability of the observations in a population and the variability of sample means from that population. If you divide the standard deviation of the observations by the square root of the number of observations, you have an estimate of the standard deviation of the sample means, also known as the standard error. So if you know what the standard error of the mean is, you can estimate what the standard deviation of the original observations must be. You just multiply the standard error by the square root of the number of cases to get an estimate of the standard deviation of the observations. You square this to get an estimate of the variance.

Using this insight, you can obtain an estimate of the variance based on *between-groups variability.* You have a sample mean for each of the groups, and you can compute how much these means vary. If the population mean is the same in all three groups, you can use the variability between the sample means (and the sizes of the sample groups) to estimate the variability of the original observations. Of course, this estimate depends on whether the population means really are the same in all three groups — which is the null hypothesis. If the null hypothesis is true, the between-groups estimate is correct. However, if the groups have different means in the population, then the between groups estimate (the estimate of variability based on the group means) will be too large.

CALCULATING THE F RATIO

You now have two estimates of the variability in the population: the within-groups mean square and the between-groups mean square. The within-groups mean square is based on how much the observations within each of the groups vary. The between-groups mean square is based on how much the group means vary among themselves. If the null hypothesis is true, the two numbers should be close to each other. If we divide one by the other, the ratio should be close to one.

The statistical test for the null hypothesis that all of the groups have the same mean in the population is based on computing such a ratio. It is called an F statistic. You take the between-groups mean square and divide it by the within-groups mean square as shown in the following formula:

F = between-groups mean square/within-groups mean square

MULTIPLE COMPARISON PROCEDURES

A significant F value tells you only that the population means are probably not all equal. It does not tell you which pairs of groups appear to have different means. You can reject the null hypothesis that all means are equal in several different situations. For example, people who find the ride comfortable may differ in age from those who find the ride rough but not from those who find the ride normal. Or people who find the ride comfortable may differ in age from both of the other groups. In

most situations, you want to pinpoint exactly where the differences are. To do this, you must use *multiple comparison procedures.*

Why do you need yet another statistical technique? Why not just calculate t tests for all possible pairs of means? The reason for not using many t tests is that when you make a lot of comparisons involving the same means, the probability that one out of the bunch will turn out to be statistically significant increases. For example, if you have five groups and you compare all pairs of means, you are making 10 comparisons. When the null hypothesis is true (that is, all of the means are equal in the population), the probability that at least one of the 10 observed significance levels will be less than .05 is about .29. If you keep looking, even unlikely events will happen. The more comparisons you make, the more likely it is that you will find one or more pairs to be statistically different, even if all means are equal in the population.

Multiple comparison procedures protect you from calling too many differences significant. They adjust for the number of comparisons you are making. The more comparisons you make, the larger the difference between pairs of means must be for a multiple comparison procedure to report a significant difference. So, you can get different results from multiple t tests than from multiple comparison procedures. Differences that the t tests find significant may not be significant based on multiple comparison procedures. When you use a multiple comparison procedure, you can be more confident that you are finding true differences.

Several different procedures can be used to make multiple comparisons. The procedures differ in how they adjust the observed significance level for the fact that many comparisons are being made. Some require larger differences between pairs of means than others. For further discussion of multiple comparisons, see Kirk (1968).

INTERACTIONS

Analysis of variance allows you to test not only for the individual variables but also for their combinations. This is an important concern. As you have noticed in previous discussions, combinations of variables sometimes have a different effect than you would expect from variables alone. That effect is important, and as experimenters we want to know about it. In statistical terms, we say that an interaction effect exists between variable X and variable Y.

ANALYSIS OF VARIANCE IN COMPUTER SOFTWARE

Each software package has its own quirks, but all of the software dealing with ANOVA has some things in common. You use the ANOVA command to perform analysis of variance with more than one factor. ANOVA does not include the multiple comparison procedures that ONEWAY offers, but it does allow you to analyze the effects and interactions of several factors at once. To analyze the effect of the main variables, you have to specify the dependent variable first and then enter the range of categories that should be used in the analysis.

All software packages show the results of an analysis in what is called the ANOVA table. This analysis of variance table may be very elaborate or very simple

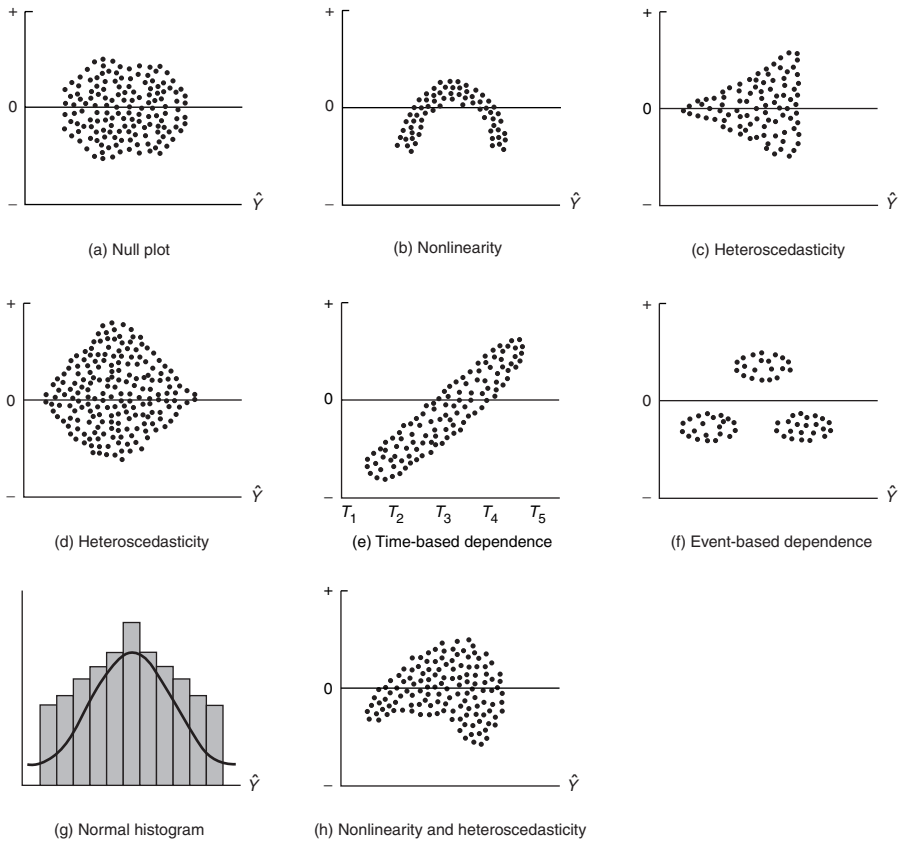


FIGURE 7.1 Typical graphical analysis of residuals.

depending on the demands that the experimenter has set for the analysis. A typical table will have columns for sums of squares, degrees of freedom, mean squares, the F ratio, and the significance of F. Under MAIN EFFECTS are the statistics for the variables under study, considered separately. (The statistics across from MAIN EFFECTS let you evaluate the significance of all the single-variable effects considered together, if you want to do that.) Under 2-WAY INTERACTIONS are the statistics for the interactions between the selected variables. The row labeled RESIDUAL contains the within-cell sum of squares and mean square. What we want to do in the residual analysis is to identify those specific variables that violate the assumption of linearity (if indeed such variables exist) and apply the needed remedies only to them. Also, the identification of outliers or influential observations is facilitated on the basis of one independent variable at a time. Typical patterns of residual analysis are shown in Figure 7.1.

For each effect in the table, the F statistic is calculated as the ratio of the mean square for that effect to the mean square for the residual.

To obtain multiple comparison tests, we enter a slash, the RANGE subcommand, and the name of the test after the previous specification. Some of the tests are:

- Least significant difference
- Duncan's multiple range test (one of the most common used)
- Student-Newman-Keul's test
- Tukey
- Tukey's honestly significant difference
- Modified least significant difference
- Scheffe's test

For a detailed explanation for each one of these, see Winer (1971).

How can you test the null hypothesis that several population means are equal? Here is what you need to know:

- Analysis of variance can be used to test the null hypothesis that several population means are equal.
- To use analysis of variance, your groups must be random samples from normal populations with the same variance.
- In analysis of variance, the observed variability in the samples is subdivided into two parts — variability of the observations within a group about the group mean (within-groups variation) and variability of the group means (between-groups variation).
- The F statistic is calculated as the ratio of the between-groups estimate of variance to the within-groups estimate of variance.
- The analysis of variance F test does not pinpoint which means are significantly different from each other,
- Multiple comparison procedures, which protect you against calling too many differences significant, are used to identify pairs of means that appear to be different from each other.

REFERENCES

Kirk, R., *Experimental Design: Procedures for the Behavioral Sciences,* Brooks, Belmont, CA, 1968.

Winer, B.J., *Statistical Principles in Experimental Design,* McGraw-Hill, New York, 1971.

8 Measuring Association

This chapter will consider how to measure the strength and nature of the relationship between two variables that have a limited number of distinct categories. What if we want to examine the relationship between two variables that are measured on an interval or ratio scale? That is what the remainder of the book is about.

OVERVIEW

One of the most frequently asked questions in any study is, "Are these two variables related?" Is education related to voting behavior? Is marital status related to happiness? Is ability to close a sale related to the experience of the salesperson? Is a particular training method related to better results? You usually want to know more than just whether the two variables are related. You also want to know the strength and nature of the relationship. If job satisfaction is related to perceiving life as exciting, how strongly is it related? And does the likelihood of perceiving life as exciting increase or decrease as job satisfaction increases? Is customer satisfaction related to loyalty? If so, how much and to what extent? Is a new gadget in a process related to increased productivity? If so, to what extent or degree?

THE STRENGTH OF A RELATIONSHIP

Many different statistical techniques are used to study the relationships among variables. We will consider some of them in the chapters that follow. In this chapter, we will look at techniques that are useful for measuring the strength and nature of associations when the two variables are categorical. These variables have a limited number of possible values, and their distribution can be examined with a cross-tabulation table.

WHY NOT CHI-SQUARE?

Previously we used the chi-square test to test the null hypothesis that two categorical variables are independent. If you reject the null hypothesis of independence, what can you say about the two variables? Can you conclude anything about the strength or nature of their association on the basis of the actual chi-square value? Do large chi-square values indicate strong associations and small values indicate weak ones?

The actual value of the chi-square statistic provides you with little information about the strength and type of association between two variables. In our discussion about the cross-tabulation, we implied that sample size will influence the results and, as a consequence, the chi-square value. If you take a particular cross-tabulation and multiply all cell frequencies by 10, you also increase the value of the chi-square by 10. By increasing the frequency in each cell, you are not in any way changing the nature or strength of the association — that remains exactly the same. The value of the chi-square statistic depends on the sample size as well as the amount of departure from independence for the two variables. So you cannot compare chi-square values from several studies with different sample sizes. This is one reason why the chi-square statistic is not very useful as a measure of association. Furthermore, since chi-square is based only on expected and observed frequencies, it is possible for many different types of tables to have the same value for the chi-square statistic. Different types of relationships between two variables can result in the same chi-square value. Knowing the chi-square tells you nothing about the nature of the association.

MEASURES OF ASSOCIATION

Statistics that are used to quantify the strength and nature of the relationship between two variables in a cross-tabulation are called *measures of association.* Many different measures of association exist because "association" can be defined in many different ways. The measures differ in how they can be interpreted and in how they define perfect and intermediate levels of association. They also differ in the level of measurement required for the variables. For example, if two variables are measured on an ordinal scale, it makes sense to talk about their values increasing or decreasing together. Such a statement would be meaningless for variables measured on a nominal scale.

No single measure of association is best for all situations. To choose the best one for a particular situation, you must consider the type of data and the way you want to define association. If a certain measure has a low value for a table, this does not necessarily mean that the two variables are unrelated. It can also mean that they are not related in the way that the measure can detect. But you should not calculate a lot of measures and then report only the largest. Select the appropriate measures in advance. If you look at enough different measures, you increase your chance of finding significant associations in the sample that do not exist in the population.

MEASURES OF ASSOCIATION FOR VARIABLES

When you have variables that are measured on a nominal scale, you are limited in what you can say about their relationship. You cannot say that marital status increases as religious affiliation increases, or that automobile color decreases with increasing state of residence. You cannot say anything about the direction of the association. If the categories of the variables do not have a meaningful order, it does not make sense to say they are associated in one direction or another. All you can do is try to measure the strength of the association. Two types of measures of association are useful for nominal variables: measures based on chi-square and measures of

proportional reduction in error (called PRE measures). We will look at each of these in turn.

Measures Based on Chi-Square

We just finished discussing why the chi-square statistic is not a good measure of association. However, since its use is common in tests of independence, people have tried to construct measures of association based on it. The measures based on chi-square attempt to modify it so it is not influenced by sample size and so it falls in the range of zero to one. Without such adjustments, you cannot compare chi-square values from tables with different sample sizes and different dimensions. (In the range from zero to one, a value of zero corresponds to no association and a value of one to perfect association. Coefficients are often *normalized* to fall in this range.)

The phi coefficient — This is one of the simplest modifications of the chi-square statistic. To calculate a phi coefficient, just divide the chi-square value by the sample size and then take the square root. The formula is

$$\phi = \sqrt{\chi^2 / N}$$

The maximum value of phi depends on the size of the table. If a table has more than two rows or two columns, the phi coefficient can be greater than one — an undesirable feature.

The coefficient of contingency — This measure is always less than or equal to one. It is often abbreviated with the letter C. It is calculated from the chi-square statistic using the following formula:

$$C = \sqrt{\chi^2 / \chi^2 + N}$$

Although the value of C is always between 0 and 1, it can never get as high as 1, even for a table showing what seems to be a perfect relationship. The largest value it can have depends on the number of rows and columns in the table. For example, if you have a four-by-four table, the largest possible value of C is .87.

Cramer's V — This is a chi-square–based measure of association that can attain the value of 1 for tables of any dimension. Its formula is:

$$V = \sqrt{\chi^2 / N(k-1)}$$

where k is the smaller of the number of rows and columns. If the number of rows or columns is two, Cramer's V is identical in value to phi.

Calculating the λ (Lambda)

The lambda statistic measures how much your error rate decreases when you use additional information about a variable. It is calculated as:

$$\lambda = \text{Misclassified in situation 1} - \text{Misclassified in situation 2/Misclassified in situation 1}$$

Lambda tells you the proportion by which you can reduce your error in predicting the dependent variable if you know the independent variable. That is why it is called a proportional reduction in error measure. The largest value that lambda can be is one. A value of zero for lambda means the independent variable is of no help in predicting the dependent variable. When two variables are statistically independent, lambda is zero; but a lambda of zero does not necessarily imply statistical independence. As with all measures of association, lambda measures association in a very specific way — reduction in error when values of one variable are used to predict values of the other. If this particular type of association is absent, lambda is zero. Even when lambda is zero, other measures of association may find associations of a different kind. No measure of association is sensitive to every type of association imaginable.

Two Different Lambdas

Lambda is not a symmetric measure. Its value depends on which variable you predict from which. Suppose that instead of predicting the excitement category based on marital happiness, you tried to predict the reverse — how happy a person's marriage was, based on how exciting the person found life to be. You would get a different value for lambda. The actual statistic will be generated from the "crosstab" job that produced the cross-tabulation table. For this discussion, you must recognize two lambdas exist: (1) the asymmetric lambda — which we just covered, and (2) the symmetric lambda.

Although we just said that the lambda value is not a symmetric statistic, sometimes if you have no reason to consider one of the variables dependent and the other independent, you can compute a symmetric lambda coefficient. You predict the first variable from the second and then the second variable from the first. The symmetric lambda is calculated as the sum of the two differences divided by the total number misclassified without additional information. In other words, you just add up the numerators for the two lambdas, then add up the denominators, then divide.

Is it really possible for variables to be related and still have a lambda of zero? That does not sound right. Actually, this can happen easily, depending on the distribution of the dependent variable. For example, consider the following cross-tabulation table:

(DEP)endent variable	**(INDEP)endent variable**			
Count	1.00	2.00	3.00	Row total
1.00	19	10	1	30
2.00	20	20	20	60
3.00	1	10	19	30
Column	40	40	40	120
Total	33.3	33.3	33.3	100.0

Using the SPSS software we generate a table that looks like:

Statistic	Symmetric	With DEP Dependent	With INDEP Dependent
Lambda	.12857	.00000	.22500
Number of Missing Observations = 0			

We can see from the above information that the two variables are clearly associated, but value two of the dependent variable occurs most often in each category of the independent variable. You would predict that value whether or not you knew the independent variable. Since knowing the independent variable does not help at all, lambda equals zero. You can see that SPSS/PC+ reports a lambda of zero when DEP is the dependent variable. Remember: a measure of association is sensitive to a particular kind of association.

Measures of Association for Ordinal Variables

Lambda can be used as a measure of association for variables measured on ordinal scales as well as for variables measured on nominal scales. The computation of lambda, however, did not use the order information. Because of that, we could rearrange the order of the rows and columns in any way we wanted and not change the value of lambda at all.

Several measures of association make use of the additional information available for ordinal variables. They tell us not only about the strength of the association but the direction as well. For example, if one variable changes in the same direction as the other, then we say that the two variables have a positive relationship. If, on the other hand, the values of one variable increase while those of the other decrease, we can say the variables have a negative relationship. We cannot make statements like these about nominal variables, since the categories of the variables have no order. Values cannot increase or decrease unless they have an order.

Concordant and Discordant Pairs

Many ordinal measures of association are based on comparing pairs of cases. For example, look at the following data, which contain a listing of the values of Var1, and Var2, for three cases.

	Var1	Var2
Case 1	1	2
Case 2	2	3
Case 3	3	2

Consider the pair of cases, Case 1 and Case 2. Both Case 2 values are larger than the corresponding values in Case 1. That is, the value for Var1 is larger for Case 2 than for Case 1, and the value for Var2 is larger for Case 2 than for Case 1. Such a pair of cases is called concordant. A pair of cases is concordant if the value of each variable is larger (or each is smaller) for one case than for the other case.

A pair of cases is discordant if the value of one variable for a case is larger than the value for the other case, but the direction is reversed for the second variable. For example, Case 2 and Case 3 are a discordant pair, since the value of Var1 for Case 3 is larger than for Case 2, but the value of Var2 is larger for Case 2 than for Case 3.

When two cases have identical values on one or both variables, they are said to be tied. Five different outcomes are possible when you compare two cases. They can be concordant, discordant, tied on the first variable, tied on the second variable, or tied on both variables. When data are arranged in a cross-tabulation, it is easy to compute the number of concordant, discordant, and tied pairs, just by looking at the table and adding up cell frequencies.

If most of the pairs are concordant, the association is said to be positive. As values of one variable increase (or decrease), so do the values of the other variable. If most of the pairs are discordant, the association is negative. As values of one variable increase, those of the other tend to decrease. If concordant and discordant pairs are equally likely, we say that no association is present.

Measures Based on Concordant and Discordant Pairs

The ordinal measures of association that we will consider are all based on the difference between the number of concordant pairs (P) and the number of discordant pairs (Q), calculated for all distinct pairs of observations. Since we want our measures of association to fall within a known range for all tables we must standardize the difference, P – Q (if possible, from –1 to 1, where –1 indicates a perfect negative relationship, 1 indicates a perfect positive relationship, and 0 indicates no relationship). The measures differ in the way they attempt to standardize P – Q.

Goodman and Kruskal's Gamma

One way of standardizing the difference between the number of concordant and discordant pairs is to use Goodman and Kruskal's gamma. You calculate the difference between the number of concordant and discordant pairs, (P – Q), and then divide this difference by the sum of the number of concordant and discordant pairs (P + Q). For example, using the previous data for Var1 and Var2, we produce a gamma value of .459. What does this mean? A positive gamma tells you that there are more "like" (concordant) pairs of cases than "unlike" (discordant) pairs. A negative gamma would mean that a negative relationship exists.

The absolute value of gamma has a proportional reduction in error interpretation. What you are trying to predict is whether a pair of cases is like or unlike. In the first situation, you classify pairs as like or unlike based on the flip of a fair coin. In the second situation, you base your decision rule on whether you find more concordant or more discordant pairs. If most of the pairs are concordant you predict "like" for all pairs. If most of the pairs are discordant you predict "unlike." The absolute value of gamma (the numerical value, ignoring a minus sign if one is present) is the proportional reduction in error when the second rule is used instead of the first.

For example, if half of the pairs of cases are concordant and half are discordant, guessing randomly and classifying all cases as concordant leads to the same number of misclassified cases — one half. The value of gamma is then zero. If all the pairs

are concordant, guessing "like" will result in correct classification of all pairs. Guessing randomly will classify only half of the pairs correctly. In this situation, the value of gamma is one.

If two variables are independent, the value of gamma is zero. However, a gamma of zero does not necessarily mean independence. (If the table is two by two, though, a gamma of zero does mean that the variables are independent.)

Kendall's Tau-b

Gamma ignores all pairs of cases that involve ties. A measure that attempts to normalize P – Q by considering ties on each variable in a pair separately (but not ties on both variables) is tau-b. It is computed as

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + T_x)(P + Q + T_y)}}$$

where T_x is the number of ties involving only the first variable, and T_y is the number of ties involving only the second variable. Tau-b can have the values of +1 and –1 only for square tables. Since the denominator is complicated, there is no simple explanation in terms of proportional reduction of error. However, tau-b is a commonly used measure.

Tau-c

A measure that can attain, or nearly attain, the values of +1 and –1 for a table of any size is tau-c. It is computed as

$$\tau_c = \frac{2m(P - Q)}{N^2(m - 1)}$$

where m is the smaller of the number of rows and columns. Unfortunately, no simple proportional reduction of error interpretation is possible for tau-c either.

Somers' d

Gamma, tau-b, and tau-c are all symmetric measures. It does not matter whether one of the variables is considered dependent. The value of the statistic is the same. Somers proposed an extension of gamma in which one of the variables is considered dependent. It differs from gamma only in that the denominator is the sum of all pairs of cases that are not tied on the independent variable. (In gamma, all cases involving ties are excluded from the denominator.)

Measures Involving Interval Data

If the two variables are measured on an interval scale, you can calculate coefficients that make use of this additional information. The Pearson correlation coefficient,

discussed later, measures the strength of what is called a *linear* association. The eta (η) coefficient can be used when a dependent variable is measured on an interval scale, and the independent variable is measured on a nominal or ordinal scale. When eta is squared, it can be interpreted as the proportion of the total variance in the dependent variable that can be accounted for by knowing the values of the independent variable.

TESTING HYPOTHESES

In addition to assessing the strength and nature of a relationship, you may want to test hypotheses about the various measures of association. For example, you may want to test the null hypothesis that the value of a measure is zero in the population. This does not involve anything new. You just have to calculate the probability that you would obtain a value as large (in absolute value) as the one you observed if the value is zero in the population. All statistical software packages print as part of their output the observed significance levels for some of the measures of association that we have discussed in this chapter.

ABOUT STATISTICS FOR CROSSTABS

Here is a list of the possible tests (most of which have been discussed in this chapter) used on the STATISTICS subcommand used with the CROSSTABS command:

Chi-square
Phi for 2×2 tables,
Cramer's V for larger tables
Contingency coefficient
Lambda
Uncertainty coefficient
    Kendall's tau-b
    Kendall's tau-c
    Gamma
    Somers' d
    Eta
    Pearson's r

How can you measure the strength of the relationship between categorical variables? Here is a summary of the things that you need to know:

- Many measures of association can be used to measure the strength of the relationship between two categorical variables.
- Measures of association differ in the way they define association.
- You should select a measure to use based on the characteristics of the data and how you want to define association.

- The chi-square statistic is not a good measure of association. Its value does not tell you anything about the strength of the relationship between two variables.
- Measures of proportional reduction in error (PRE) compare the error you make when you predict values of one variable based on values of another with the error when you predict them without information about the other variable.
- Special measures of association are available for ordinal variables. They are based on counting the number of concordant pairs (as one variable increases, so does the other) and the number of discordant pairs (as one variable increases, the other decreases).

PLOTTING

So far we have discussed various statistical tests to measure the strength of a relationship between two variables. Another way to evaluate the relationships between variables is plotting. The graphical representation makes the evaluation much easier and provides a good starting point for further investigation. Reference lines on plots can make it easier to see the patterns that indicate relationships in the data.

Plotting data is one of the best ways to look for relationships and patterns. A plot is simple to understand and conveys a lot of information about the data. In other chapters, we discussed methods of summarizing and describing relationships, but those methods are no substitute for plots. Whenever possible, you should plot the data first, and then think about appropriate methods for describing the plots. We made this point earlier in our discussion about histograms.

A plot can also alert you to possible problems in the data. For example, if you are plotting salary and age, and you find a 22-year-old with a salary of $100,000, you have reason to be suspicious. Although the values may be correct, it is much more likely that either the age or the salary was recorded or entered incorrectly. You would not have been able to pick out this point as suspicious if you examined the variables individually. A salary of $100,000 is high but possible. An age of 22 is not unusual. It is the combination of the values that leads you to suspect the point may be a mistake. Even if the point is correct, it is important to identify it early, since it may need special treatment in later analyses.

With the analysis of the data and the plotting we can see if the variables under study "appear to be related." What do we mean by related here? Nothing really complicated. Two variables are related if knowing the value of one variable tells us something about the value of the other variable. Neither of the variables has to be considered dependent or independent. All we are interested in is how they behave together. Usually the behavior is shown as a straight line through the data.

Five fundamental types of relationships are possible: (a) strong positive relationship, (b) weak positive relationship, (c) no relationship, (d) strong negative relationship and (e) weak negative relationship. These relationships are shown in Figure 8.1.

In real life, linear relationships look more like a cluster of points scattered all around with a straight line through the center of the points. Figure 8.2 shows an

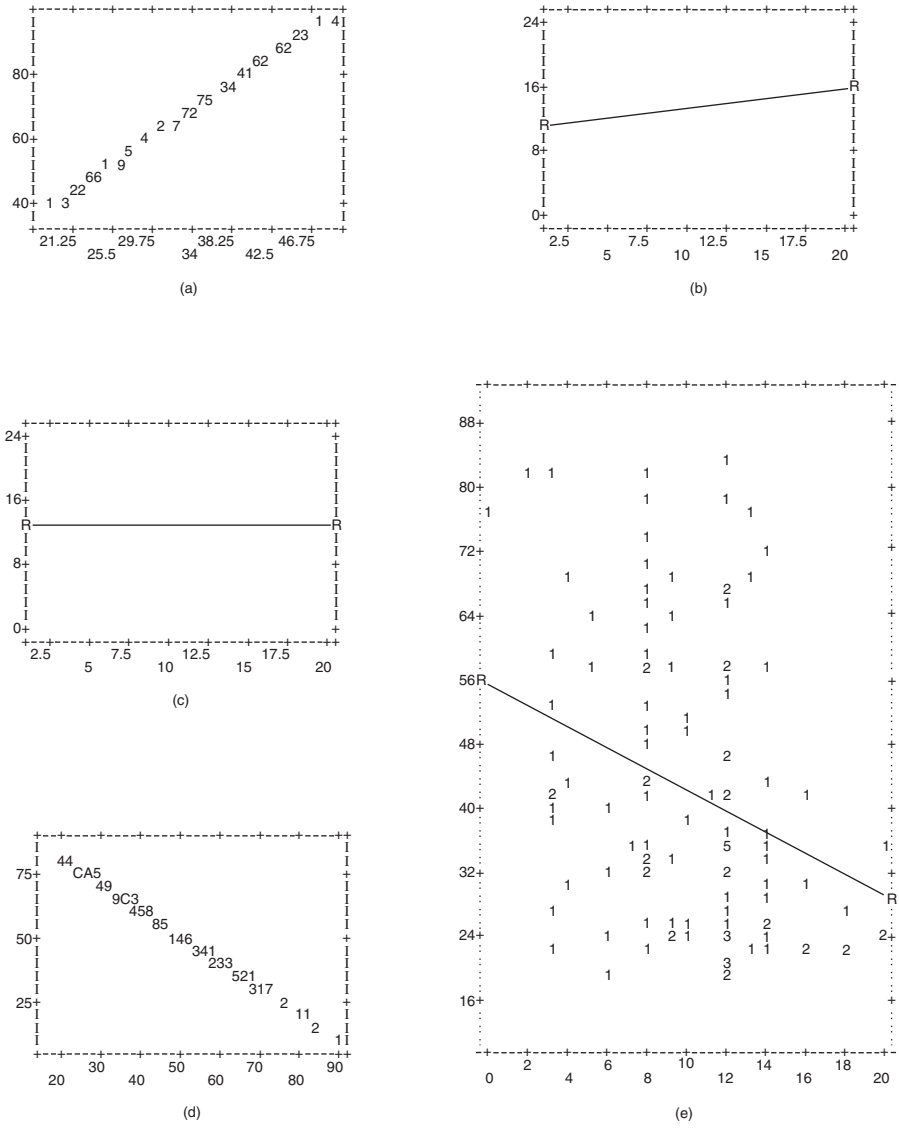


FIGURE 8.1 Five types of relationships.

example. The points are clustered around the line, but most of them do not fall exactly on it. Instead they are distributed around it. You can see quite a bit of variability around the line, but most points are not too far removed. That is OK, because you can still see that the relationship is positive, since as one variable moves in one way so does the other. Remember we are interested only in the direction and magnitude of the relationship. To find the best place to draw the line we use the regression method, which we will cover later.

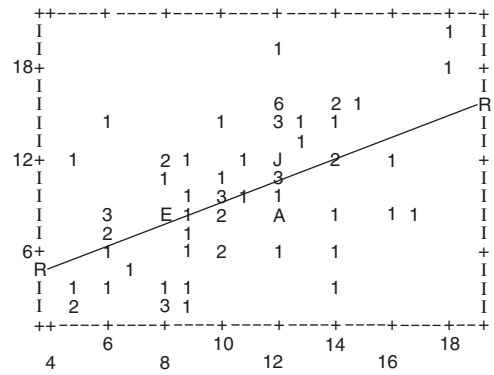


FIGURE 8.2 A relationship with points scattered around a straight line.

How can you display the relationship between two variables that are measured on an interval or ratio scale? Here is a summary:

- A plot displays the values of two variables for each case.
- By examining a plot, you can see what sort of relationship, if any, exists between two variables.
- The points on a plot may be identified by their values on an additional variable (called a control variable). This lets you see whether the relationship between the two variables differs for the different categories of the control variable.
- Points that have unusual combinations of values can be identified from a plot, since they will be far removed from the other cases.

COVARIANCE

All of the summary measures to this point involve a single variable. It is also useful to summarize the relationship between two variables. Specifically, we would like to summarize the type of behavior often observed in a scatterplot. Two such measures are *covariance* and *correlation*. We will discuss them briefly here and in more depth in later chapters. Each measures the strength (and direction) of a linear relationship between two numerical variables. Intuitively, the relationship is "strong" if the points in a scatterplot cluster tightly around some straight line. If this straight line rises from left to right, then the relationship is "positive" and the measures are positive numbers. If it falls from left to right, then the relationship is "negative" and the measures are negative numbers.

First, it is important to realize that if we want to measure the covariance or correlation between two variables X and Y — indeed, even if we just want to form a scatterplot of X vs. Y — then X and Y must be "paired" variables. That is, they must have the same number of observations, and the X and Y values for any observation should be naturally paired. For example, each observation could be the

height and weight for a particular person, the time in a store and the amount of money spent by a particular customer, and so on.

With this in mind, let X_i and Y_i be the paired values for observation i, and let n be the number of observations. Then the covariance between X and Y, denoted by Cov (X, Y), is given by the formula

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n - 1}$$

You probably will not ever have to use this formula directly — most software packages have a built-in COVAR function that does it for you — but the formula does indicate what covariance is all about. It is essentially an average of products of deviations from means. If X and Y vary in the same direction, then when X is above (or below) its mean, Y will also tend to be above (or below) its mean. In either case, the product of deviations will be positive — a positive times a positive or a negative times a negative — so the covariance is positive. The opposite is true when X and Y vary in opposite directions. Then, the covariance is negative.

The limitation of covariance as a descriptive measure is that the units in which X and Y are measured affect it. For example, we can inflate the covariance by a factor of 1000 simply by measuring X in dollars rather than in thousands of dollars. To remedy this problem the correlation is used.

When there are more than two variables in a data set, it is often useful to create a table of covariances and/or correlations. Each value in the table then corresponds to a particular pair of variables.

CORRELATION

Although plots give you a pretty good idea of the strength of a linear association, they do not provide an objective summary measure that you could use to compare and summarize the relationships between pairs of variables. On the basis of the plots, you could say that a relationship between variable X and variable Y appears to be present (it may be strong or weak depending on the slope of the line), but you cannot really say how strong the relationship is unless you have a summary measure that quantifies your visual impressions. This is reflected in Figure 8.3. In this figure a combination of outcomes is summarized in matrix form. Values above the diagonal are bivariate correlations, with corresponding scatterplots below the diagonal. The diagonal portrays the distribution of each variable.

The most commonly used measure is the *Pearson correlation coefficient,* which is abbreviated as r. (The statistic is named after Karl Pearson, an eminent statistician of the early twentieth century.) Here are some of its characteristics:

- If no linear relationship exists between two variables, the value of the coefficient is 0.

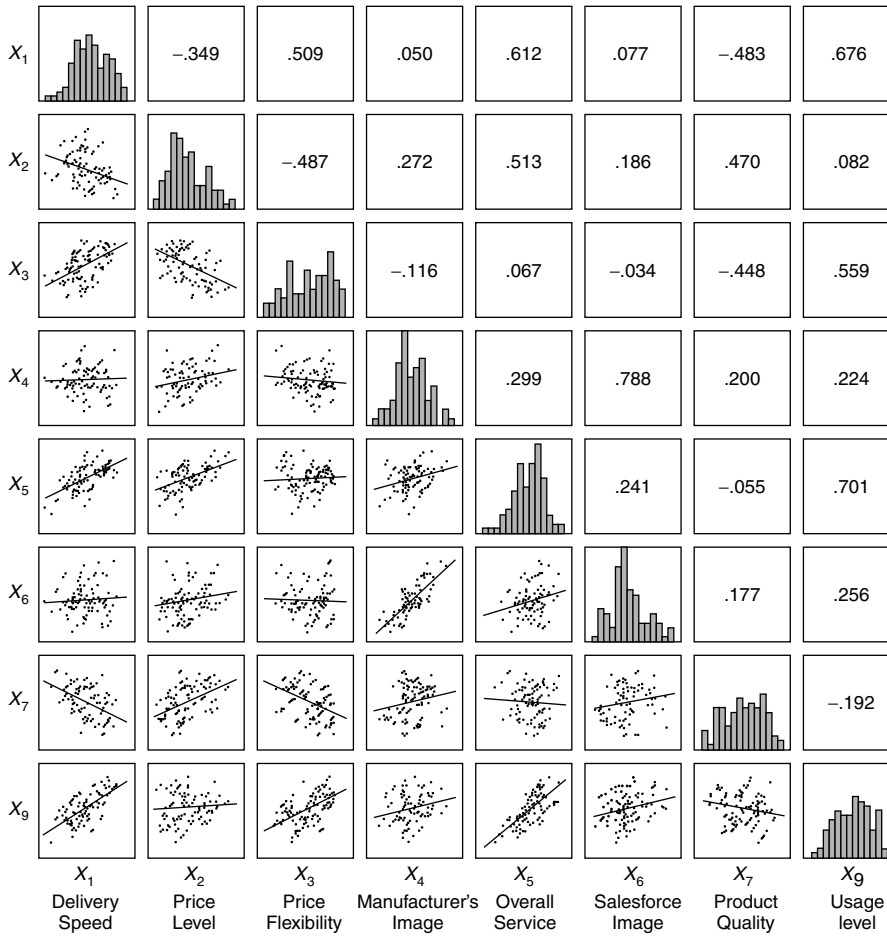


FIGURE 8.3 Scatterplot matrix of metric variables.

- If a perfect positive linear relationship is present, the value is +1.
- If a perfect negative linear relationship is present, the value is –1.

To summarize, the values of the coefficient can range from –1 to +1, with a value of 0 indicating no linear relationship. Positive values mean that a positive relationship exists between the variables. Negative values mean that a negative relationship is present. If one pair of variables has a correlation coefficient of +.8, while another pair has a coefficient of –.8, the strength of the relationship is the same for both. The direction of the relationship differs, however.

Does a correlation coefficient of zero mean that no relationship exists between two variables? Not necessarily. The Pearson correlation coefficient only measures the strength of a linear relationship. Two variables can have a correlation coefficient close to zero and yet have a very strong nonlinear relationship. Look at Figure 8.4, which is a plot of two hypothetical variables. You will note that a strong relationship

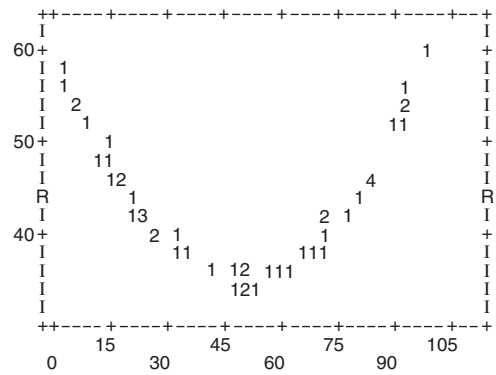


FIGURE 8.4 Strong relationship but very low correlation.

exists between the two variables. The value of the correlation coefficient, however, is close to zero. Always plot the values of the variables before you compute a correlation coefficient. This will allow you to detect nonlinear relationships, for which the Pearson correlation coefficient is not a good summary measure. The Pearson correlation coefficient should only be used for linear relationships.

The mathematical formula that tells you how to calculate the correlation coefficient for a pair of variables is

$$r = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{(N-1)S_x S_y}$$

where X and Y are the values of the two variables for a case, N is the number of cases, and S_x and S_y are the standard deviations of the two variables. It does not matter which variable you take to be X and which to be Y in the formula, since the correlation coefficient will be the same. The correlation coefficient is not expressed in any unit of measurement. The correlation coefficient between two variables will be the same regardless of how you measure them.

Sometimes the correlation coefficient is used simply to summarize the strength of a linear relationship between two variables. In other situations you may want to do more than that; you may want to test hypotheses about the population correlation coefficient. For example, you may want to test the null hypothesis that no linear relationship exists between variable X and variable Y in the population. Remember, if your data are a random sample from a particular population, you want to be able to draw conclusions about the population based on the results you observe in your sample. As was the case with other descriptive measures such as the mean, you know that the value of the correlation coefficient you calculate for your sample will not exactly equal the value that you would obtain if you had values for the entire population. You know that if you took many samples from the same population and

calculated the correlation coefficients, their values would vary. That is, there is a distribution of possible values of the correlation coefficient, just as there is a distribution of possible values for sample means. If you know what the distribution is, you can calculate observed significance levels. For example, you can calculate how often you would expect to find, in samples of a particular size, a coefficient of .3 or greater when the population value is zero.

Does Significant Mean Important?

If you reject the null hypothesis, does that mean that an important relationship exists between the two variables? No. It simply means that it is unlikely that the value of the correlation coefficient is zero in the population. For large sample sizes, even very small correlation coefficients have small observed significance levels. You can have a correlation coefficient of .1 and have it be statistically significant. It indicates that a very small, but nonzero, linear relationship exists between the variables. You should look at both the value of the coefficient and its associated significance level when evaluating the relationships among variables.

One-Tailed and Two-Tailed Significance Probabilities

If you do not know before looking at your data whether a pair of variables should be positively or negatively correlated, you must use a two-tailed significance level. You reject the null hypothesis for either large positive or large negative values of the correlation coefficient. If you know in advance whether your variables should be positively or negatively correlated, you can use a one-tailed significance test, For example, if you are studying the relationship between total yearly income and value of housing, you know that if a relationship exists, it will be positive. Poor people cannot own expensive houses.

For a one-tailed test, you reject the null hypothesis only if the value of the correlation coefficient is large and in the direction you specified. For a one-tailed test, the observed significance level is half of the two-tailed value. That is because you only calculate the probability that you would obtain a more extreme value in one direction, not two. If you do not specify what kind of test you want when you use a software package, more often than not the default is a one-tailed test. In order to get two-tailed tests, you must specify it with specific option of the software.

Assumptions about the Data

In order to test hypotheses about the Pearson correlation coefficient you have to make certain assumptions about the data. If your data are a random sample from a population in which the distribution of the two variables together is normal, the previously described procedure is appropriate. If it seems unreasonable to assume that the variables are from normal distributions, you may have to use other statistical procedures that do not require the normality assumption. These are nonparametric procedures and are described in detail in statistics books such as Siegel (1956). Some of them will be summarized at the end of this chapter.

Examining Many Coefficients

If your study involves many variables, you may be tempted to compute all possible correlation coefficients among them. If you are just interested in exploring possible associations among the variables, you may find the coefficients helpful in identifying possible relationships. However, you must be careful when examining the significance levels from large tables. If you have enough coefficients, you expect some of them to be statistically significant even if no relationship exists between the variables in the population. If you compute 100 coefficients, you expect somewhere around five (95% confidence) of them to have observed significance levels less than .05 even if none of them are truly related. Think about it — that is what a significance level means.

When dealing with large samples and many variables, we must be concerned about missing values. If our study is complete and all the values have been accounted for, then we call this process *pairwise deletion* of missing data. Our calculation used as much of the data as possible. On the other hand, analysis of data when some of the cases have missing information can be troublesome, especially if you have reason to believe that the missing values are related to values of one of the variables you are analyzing.

If your data have any missing values, you should see whether the missing values show a pattern. For example, you can calculate the average of the variables. If these values are quite different, you have reason to suspect that the data values are not randomly missing. When values are not randomly missing, you must use great caution in attempting to analyze the data. In fact, you may not be able to analyze some of it. (With pairwise deletion of missing data, you could even end up with correlation coefficients that were based on entirely different groups of cases.)

Perhaps one of the most confusing issues in correlation is the notion of whether correlation and cause are the same. So, if two variables are correlated, does that mean one of them causes the other? Not at all. You can never assume that just because two variables are correlated, one of them causes the other. If you find a large correlation coefficient between the ounces of coffee consumed in a day and number of auto accidents in a year, you cannot conclude that coffee consumption causes auto accidents. It may well be that coffee drinkers also consume more alcohol, or are older, or are more poorly coordinated than people who do not drink coffee. You cannot easily tell which of the factors may influence the occurrence of accidents.

How can you summarize the strength of the linear relationship between two variables? Here are some key points to remember:

- The Pearson correlation coefficient measures the strength of the linear relationship between variables.
- Two variables have a positive relationship if, as the values of one variable increase, so do the values of the other.
- Two variables have a negative relationship if, as the values of one variable increase, the values of the other decrease.
- A correlation coefficient of +1 means that a perfect positive linear relationship exists between two variables. A value of –1 means that a perfect negative linear relationship exists.

- A correlation coefficient only measures the strength of a *linear relationship.* If a strong nonlinear relationship between two variables is present, the correlation coefficient can be zero.
- A correlation between two variables does not necessarily mean that one causes the other.
- To test the null hypothesis that the correlation coefficient is zero in the population, you can calculate the observed significance level for the coefficient.
- You can use a one-tailed test if you know in advance whether the relationship between two variables is positive or negative.
- If you have missing values in your data, you should see whether there is a pattern to the cases for which information is missing.

REFERENCE

Siegel, S., *Nonparametric Statistics for the Behavioral Sciences,* McGraw-Hill, New York, 1956.

9 Calculating Regression Lines

The last chapter discussed the measurement of associations between variables and focused on the linearity of that relationship. In this chapter we will define that relationship and its strength through the mathematics of the straight line.

OVERVIEW

As you saw in the last chapter, the correlation coefficient provides you with a measure of the strength of the linear association between two variables. All it measures, though, is how closely the points cluster about a straight line. It does not tell you anything about the line itself. In many situations, it is useful to obtain information about the actual line that is drawn through the data points. For example, if a linear relationship exists between two variables, and you know the equation for the line that describes their relationship, you can use one variable to predict values of the other. If a linear relationship exists between the amount of money a company spends on advertising and its sales volume for the year, you can use the line to predict sales volume based on advertising expenditures. That is what this chapter is about. You will learn how to calculate what is called a regression line and what it means.

CHOOSING THE BEST LINE

Thinking back to the last chapter, you will remember that the correlation coefficient was based on how closely points cluster about "a line." We did not say anything about how we selected the line shown on the plots. We will consider that now. When the correlation coefficient is +1 or –1, all the data points fall on a single line. All you have to do is connect the points and you have a line. You do not have to worry about choosing a line. When the observations are not perfectly correlated, many different lines may be drawn through the data. How do we choose among them? Since we want a line that describes the data, it should be as close as possible to the points. "As close as possible" can be defined in different ways. The most commonly used method for determining the line is called the method of least squares. The *least squares line* is the line that has the smallest sum of squared vertical distances from the observed points to the line.

THE EQUATION OF A LINE

Before we talk any more about lines, we need to consider the equation for a straight line. Take two variables, say X and Y. Call the variable we plot on the vertical axis Y, and the variable we plot on the horizontal axis X. Then the equation is

$$Y = A + B \times X$$

The A and B in this equation are just numbers. The value A is called the *intercept*, and the value B is called the *slope*. The slope is the angle of the line and the intercept is the point where it crosses the vertical axis. If the value for the slope is positive, it tells you that as one variable increases, so does the other variable. If the slope is negative, it tells you that as one variable increases, the other decreases. If the slope is large, the line is steep, indicating that a small change in one of the variables would lead to a large change in the other variable. If the slope is small, the increase or decrease is gradual. If the slope is zero, it means that changes in the X variable have no effect on the Y variable.

PREDICTING VALUES FROM THE REGRESSION LINE

The equation for the least squares regression line for a dependent variable Y and an independent variable X is

$$\text{Predicted } Y = 10.24 + .27 \times X$$

where 10.24 is the intercept, .27 is the slope of the X, and of course the predicted Y is the value expected based on the calculated numbers. (The numerical numbers are used here for the purpose of explaining the format of the equation. In real terms, these are the numbers that we are trying to identify with the regression analysis.)

Once you have a regression equation, it is easy to obtain predicted values. All you have to do is substitute the value of the independent variable into the equation.

CHOOSING THE DEPENDENT VARIABLE

When we are computing correlation coefficients it does not matter which variable we plot on the horizontal axis and which we plot on the vertical. The correlation coefficient is exactly the same. That is not usually true for regression. The slope and the intercept will differ depending on which variable is the Y variable in the equation and which is the X variable. Since regression analysis is used to predict values of a dependent variable from values of an independent variable, Y is taken to be the dependent variable and X the independent.

In this example, X is the independent variable and Y is the dependent variable. Although the regression line is a useful summary of the relationship between two variables, the values of the slope and intercept alone do not indicate how well the line actually fits the data. We need some measure of goodness of fit. We know that if the regression line fits the data perfectly, the observed values for the dependent

variable equal the predicted values. They all fall exactly on the line. The more poorly the line fits, the more discrepancy we would expect between the line and the actual values.

CORRELATING PREDICTED AND OBSERVED VALUES

One way we can measure how well the line fits is to calculate a correlation coefficient between the observed values of the dependent variable and those predicted from the regression equation. The value will be one if a perfect fit exists and close to zero if the fit is poor.

The correlation coefficient between the observed and predicted values is exactly the same value we obtained for the correlation of the two variables by using the method of the last chapter. So, now we have yet another interpretation of the correlation coefficient. It is a measure of the strength of the linear relationship between the observed values of the dependent variable and those predicted by the regression line. The correlation coefficient tells us how well the least squares line fits the data. This interpretation applies only to the absolute value of the correlation coefficient (its value if you disregard the sign). That is because even if the relationship between two variables is negative, the relationship between the observed and predicted values will be positive.

If you square the value of the correlation coefficient, you obtain yet another useful statistic. The square of the correlation coefficient tells you what proportion of the variability in the dependent variable is "explained" by the regression. What do we mean when we say that the regression "explains" variability? In general, the distance between a point and the regression line is a measure of how much variability we cannot explain with the regression line. If you compare the sum of the squared distances from the data points to the regression line with the total variability in the dependent variable, you can calculate what percent of the total variability is unexplained by the regression. The remainder of the variability is explained. This is what the square of the correlation coefficient tells you. It is the proportion of the total variability in the dependent variable that can be accounted for by the independent variable.

THE POPULATION REGRESSION LINE

When you calculate a regression line that is used only to *describe* an observed relationship between two variables, you have two concerns. Are the variables measured on an interval or ratio scale, and does their relationship appear to be linear? It makes no sense to calculate a regression line relating religious preference to the region in which someone lives. The categories of these variables have no order, and a statistic such as the slope is meaningless. Even if the two variables are measured on an interval scale, it makes no sense to calculate a regression line if their relationship is not linear. You may need to fit some other mathematical function besides a straight line, or perhaps change the scale on which the variables are measured. Those topics are mostly beyond the scope of this book, but we will talk a little about them in our discussion about residuals.

When you are interested in drawing conclusions about the *population* regression line, you need additional assumptions. First, we need to clarify what we mean by a "population regression line." In all of our previous discussions about hypothesis testing, we considered our data to be a random sample from some underlying population. We wanted to draw conclusions about the population based on what we saw in our sample. When we computed a sample mean, we considered it to be our best guess of the population mean. When we computed a correlation coefficient, we considered it our best guess for the value of the correlation coefficient in the population.

What we will be doing now is very similar. We will try to draw conclusions about the relationship of two variables in the population based on the results we see in our sample. If we had been able to include our entire population in the study, we could calculate a regression line that describes the relationship between the two variables in the population. This would be the "true" or *population regression line.* We do not know what the true line is, since all we have is a sample from the population. We do not know the true slope or the true intercept. We do have some evidence about what they are, however. Our best guess for the population line is the results observed in our sample.

To be able to test hypotheses about the population line statistically, we must make some assumptions about the population. We need these assumptions so we will know that the sampling distributions of the slope and intercept will be normal. (The sampling distribution of the slope is the distribution of the values of the slope that you would get if you took all possible samples of a particular size from a population. The sampling distribution of the intercept is defined similarly.) As before, our computations of the observed significance level will be based on these sampling distributions.

To be able to test a hypothesis about the population line statistically, we must make some additional assumptions about the population. They are:

1. The distribution is normal.
2. All of these distributions have the same variance.
3. Linearity exists.
4. All observations are selected independently.

We have just stated that to test hypotheses about the population line, we need to assume that the distributions of the dependent variable must be normal for each value of the independent variable and that the variances of those distributions must be equal. Now, what about the means of the distributions? If a linear relationship exists in the population between the two variables, then the means of all of the population distributions must fall on a straight line. To test regression hypotheses, we must assume that this is true. Look at Figure 9.1, which schematically shows the assumptions we have been talking about. In the population, a true regression line exists that specifies the relationship between the variables. This line is drawn in on the plot. For each value of the independent variable there is a distribution of the values of the dependent variable. These distributions are all normal and have the same variance. The means of all of these distributions fall on a straight line.

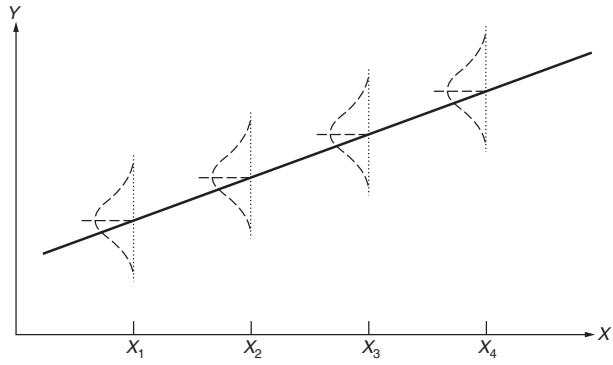


FIGURE 9.1 Regression assumptions.

The last assumption that we need for linear regression analysis is that all observations are selected independently. That is, including one observation/person in the sample should not in any way alter the chance of any other observation/person being included.

SOME HYPOTHESES OF INTEREST

We will return to the question of how you can examine your data to see if they violate any of these assumptions, but meanwhile, suppose that the data do satisfy all of the assumptions outlined above. What sorts of hypotheses can we now test? We can test hypotheses about the values of the population slope and the population intercept and hypotheses about how well the regression model fits the population. With the computer, this is no problem. However, one has to watch out for the coefficients of the regression, because instead of thinking of these values as descriptions of the sample, we will now treat them as our best guesses, or estimates, of the unknown population values of the slope and intercept. Another thing you have to watch is the fact that the coefficients and plot values may not be exactly the same, even though they are the same variables. This is because they are not all the same cases. Previously we took a smaller sample so we would not have too many cases to plot. We would expect the slope and intercept in a plot to be better estimates of the population values than the values that we are considering now, since these are now based on more cases.

It is also important to recognize that since both the slope and the intercept are calculated numbers there is variation within them. That variation is identified as the standard error of the slope and intercept. The standard errors are estimates of the standard deviation of the sampling distributions of the slope and the intercept. Remember, the slope and intercept we have calculated are based on one sample from a population. If you took another sample and calculated values for the slope and intercept they would differ. The values of the slope and intercept from repeated samples from the same population have a distribution. The standard deviation of this distribution is called the standard error. It is just like the standard error of the mean.

ARE THE POPULATION VALUES ZERO?

If no linear relationship exists between two variables in the population, the true slope is zero. All of the means of the distributions are the same. Even if the population value is zero, of course, you would not expect the sample value for the slope to be exactly zero. You hope that it would not be too far from zero, though. To test the null hypothesis that the value of the slope is zero in the population, we can calculate the probability of obtaining a slope at least as large as the one we have observed when the null hypothesis is true. As usual, if this probability is small we will reject the null hypothesis that the slope is zero. The observed significance level is based on the *t* statistic. This *t* statistic is calculated (like any *t* statistic) by dividing a sample value by its standard error. In this case, the *t* values are the sample slope and the sample intercept, divided by their standard errors. In association with this two-tailed *t* test, there is also a reported significance level for the tests of the hypotheses that the slope and intercept are zero in the population.

When testing whether a linear relationship exists between variables, the important test is the test of the slope. The intercept is simply the value of the dependent variable when the independent variable is zero. All that the test of the intercept tells us is whether the regression line goes through the origin. (The origin is the point at the intersection of the two axes. It is the point where both variables are zero.)

CONFIDENCE INTERVALS FOR REGRESSION COEFFICIENTS

The sample values for the slope and intercept are our best guesses for the population values. However, we know it is unlikely that they are exactly on target. As we have discussed before, it is possible to calculate a confidence interval for the population value. A confidence interval is a range of values that, with a designated likelihood, contains the unknown population value. To obtain 95% confidence intervals for the slope and intercept using for example the SPSS/PC software we identify the command REGRESSION, but we must add an additional specification to the command. It is called the STATISTICS subcommand, and it tells the system what values we want to see printed. That is all. The output is going to give us not only the line of the regression but also the confidence intervals.

Remember what 95% confidence means: if we draw repeated samples from a population, under the same conditions, and compute 95% confidence intervals for the slope and intercept, 95% of these intervals should include the unknown population values for the slope and intercept. Of course, since the true population values are not known, it is not possible to tell whether any particular interval contains the population values. Quite often, neither the confidence interval for the slope nor the one for the intercept contains the value zero. An interval will only include zero if you cannot reject the null hypothesis that the slope or intercept is zero, at an observed significance level of .05 or less.

GOODNESS OF FIT OF THE MODEL

We already have discussed the importance of assessing how well the regression model actually fits the data. The REGRESSION command prints several statistics that describe the "goodness of fit." Here are some of the most common ones:

- MULTIPLE R is just the absolute value of the correlation coefficient between the dependent variable and the single independent variable. It is also the correlation coefficient between the values predicted by the regression model and the actual observed values. If the value is close to one, the regression model fits the data well. If the value is close to zero, the regression model does not fit well.
- Another way of looking at how well the regression model fits is to see what proportion of the total variability (or variance) in the dependent variable can be "explained" by the independent variable. The variability in the dependent variable is divided into two components: variability explained by the regression, and variability not explained by the regression. Because of the way they are calculated, these two components are termed *sums of squares*. Indeed, they are conceptually very similar to the sums of squares we mentioned at the end of Chapter 7. The sums of squares explained by the regression equation are labeled REGRESSION, while the unexplained variability is labeled RESIDUAL. You can obtain the total variability in the dependent variable by adding up these two sums of squares. To calculate what proportion of the total variability is explained by the regression, all you have to do is divide the regression sum of squares by the total sum of squares.
- You can calculate this proportion in an easier way. All you have to do is square the correlation coefficient. This value is called R SQUARE. From R^2 we see that in the sample we can explain X% of the variability in one variable by knowing something about the other variable.
- Yet another way to test the null hypothesis that no linear relationship exists between the two variables is analysis of variance. The actual test is the F test. F is the ratio of the mean square for regression to the mean square for the residual, and the mean squares are the sums of squares divided by their respective degrees of freedom. You can find these mean squares by looking at the output of the regression output, identified as *Mean Square.* If no linear relationship exists between the two variables, then each of these mean squares provides an estimate of the variance, or variability, of the dependent variable. If a linear relationship exists, then the variability estimate based on the regression mean square will be much larger than the estimate of variability based on the residuals. Large F values suggest that a linear relationship exists between the two variables.

Is this F statistic related to the test that the slope is zero? It seems as though we are testing the same hypothesis in both situations. Yes, the two tests are evaluating exactly the same hypothesis when only one independent variable exists. In fact, a relationship exists between the two statistics. If you square the *t* value for the test that the slope is zero, you will come up with the F value in the analysis of variance table. (Try it.) For a simple equation such as this, you do not learn anything from the analysis of variance table that you did not already know from the test of the slope.

The STANDARD ERROR is yet another statistic that is available when you run a regression analysis. It is an estimate of the standard deviation of the distributions

of the dependent variable. Remember, we assumed that for each value of the independent variable there is a distribution of values of the dependent variable. All of these distributions are normal and have the same standard deviation. Our estimate of it is the standard error.

Another powerful option is the ADJUSTED R SQUARE, which is most useful when you have a model with several independent variables. This statistic adjusts the value of R^2 to take into account the fact that a regression model always fits the particular data from which it was developed better than it will fit the population. When only one independent variable is present and the number of cases is reasonably large, the adjusted R^2 will be very close to the unadjusted value. (This statistic is very useful when we check for measurement error.)

MULTIPLE REGRESSION

The REGRESSION command can use more than one independent variable in the same equation ("multiple regression"). To use multiple independent variables, just name them all on the VARIABLES subcommand. All the variables except the one you name on the DEPENDENT subcommand are used in the equation.

With multiple independent variables, you can specify how they should be entered into the equation using the METHOD subcommand. The following methods are available.

ENTER: *Enter a group of variables all at once.* This is the method we have been using throughout the chapter. You can specify the names of specific variables after the keyword ENTER.

REMOVE: *Remove a group of variables all at once.* These must be variables that entered the equation on a previous METHOD subcommand. You must specify the names of specific variables after the keyword REMOVE.

FORWARD: *Enter the variables one at a time.*

BACKWARD: *Remove the variables one at a time.* If some variables are already in the equation from a previous METHOD subcommand, they are removed one at a time. Otherwise, all variables are entered and then removed one at a time.

STEPWISE: *Enter and remove variables one at a time,* until the F statistics do not indicate that any variables in the equation should be removed or that any variables that are on the VARIABLES subcommand but not in the equation need to be entered.

You can specify several METHOD subcommands after a single DEPENDENT subcommand. The methods are applied one after the other.

How can you test hypotheses about the population regression line, based on the values you obtain in a sample? Here are the important principles:

- To draw conclusions about the population regression line, you must assume that for each value of the independent variable, the distribution of values of the dependent variable is normal, with the same variance. The means of these distributions must all fall on a straight line.

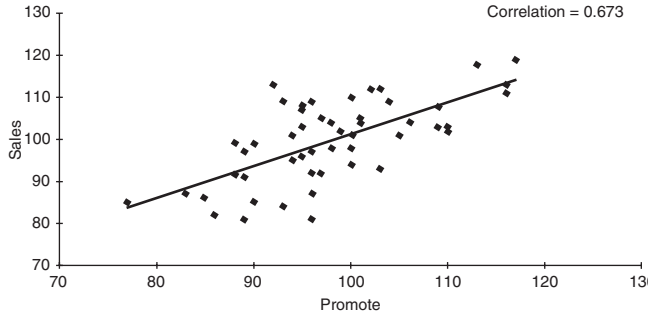


FIGURE 9.2 Scatterplot with possible linear fit superimposed.

- The test of the null hypothesis that the slope is zero is a test of whether a linear relationship exists between the two variables.
- The confidence interval for the population slope provides you with a range of values that, with a designated likelihood, includes the population value.
- When a single independent variable is present, the analysis of variance table for the regression is equivalent to the test that the slope is zero.

RESIDUALS

When you begin studying the relationship between two variables you usually do not know whether the assumptions needed for regression analysis are satisfied. You do not know whether a linear relationship exists between the two variables, much less whether the distribution of the dependent variable is normal and has the same variance for all values of the independent variable. One of the goals of regression analysis is to check whether the required assumptions of linearity, normality, and constant variance are met. To do this we do an analysis of residuals.

A quantity called the *residual* plays a very important role when you are fitting models to data. You can think of a residual as what is left over after a model is fit. In a linear regression, the residual is the difference between the observed and predicted values of the dependent variable. If a person has 12 years of education and your model predicts nine, the residual for the case is 12 – 9 = 3. You have three years of education left over (not explained by the model).

By looking at the residual for each case you can see how well a model fits. If a model fits the data perfectly, all of the residuals are zero. Cases for which the model does not fit well have large residuals. Obviously, you can use the REGRESSION procedure to calculate the residuals for all of the cases. A typical scatterplot with a linear fit is shown in Figure 9.2 and a plot of fitted values and residuals in Figure 9.3.

JUDGING THE SIZE OF THE RESIDUALS

How can you tell whether a residual is big or small? If I tell you that a case has a residual of 500, can you say whether the model gives a reasonably good prediction

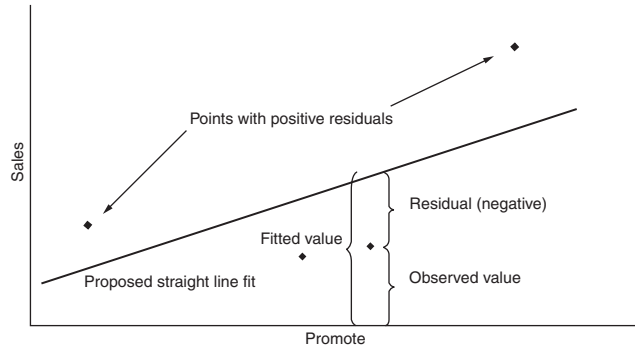


FIGURE 9.3 Fitted values and residuals.

for the case? On first thought, 500 seems to be a pretty large number — an indication that a model does not fit that case. However, if you are predicting income in dollars, a residual of 500 may not be all that large. Predicting a person's income to the nearest 500 dollars is pretty good. On the other hand, if you're predicting years of education, a residual of 500 should send you searching for a new and improved model. One way to modify the residuals so that they would be easier to interpret is to standardize them. That is, divide each residual by an estimate of its standard deviation.

How come you are only dividing the residual by its standard deviation? Why you are you not first subtracting off the mean, as you did before, when computing standardized values? You do not need to subtract the mean of the residuals before dividing by the standard deviation because the mean of the residuals is zero. If you add up all of the residuals, you will find that their sum, and therefore their mean, is zero. That is always true for a regression model that includes a constant.

For most cases, the standardized residuals range in value from –2 to +2 and they are identified as *ZRESID. (Remember that in a normal distribution with a mean of 0 and a standard deviation of 1, about 95 percent of the cases fall within +2 and –2.) Whenever you see a standardized residual larger than +2 or smaller than –2, you should examine the case to see if you can find some explanation for why the model does not fit. A typical presentation of residuals in a standardized format is shown in Figure 9.4.

LOOKING FOR OUTLIERS

If you have a large number of cases, you may not want to look at the values of the residuals for all of them. Instead you may want to look only at the cases with "large" residuals. Such cases are called *outliers*. This is easy to do with the REGRESSION procedure. Just leave off the keyword ALL on the CASEWISE subcommand. In most computer software applications, you will find that if you do not tell the program to print all cases, it prints only those whose standardized residuals are greater than 3 or less than –3. By looking at the characteristics of the outliers, you can see situations where the model does not work well. Typical outliers

will appear with large residuals in the form of Figure 9.5. The effect on the actual regression line is shown in Figure 9.6 and finally the outliers outside the pattern are shown in Figure 9.7.

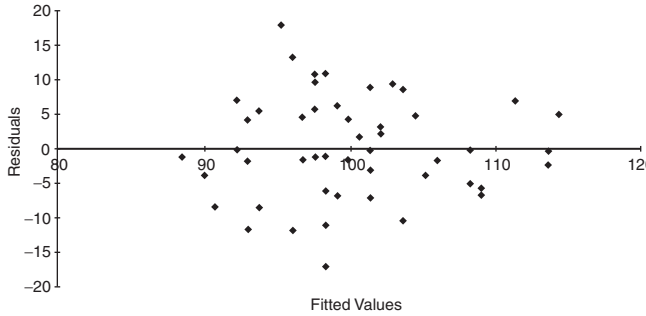


FIGURE 9.4 Typical residuals in a standardized format.

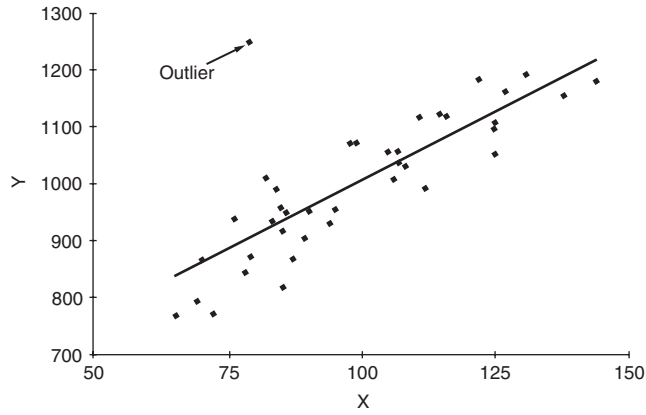


FIGURE 9.5 Outlier with large residual.

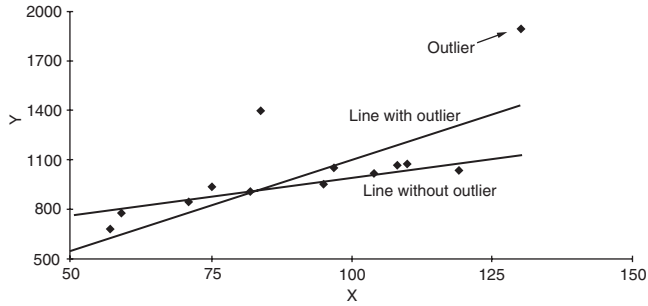


FIGURE 9.6 Outlier that tilts the regression line.

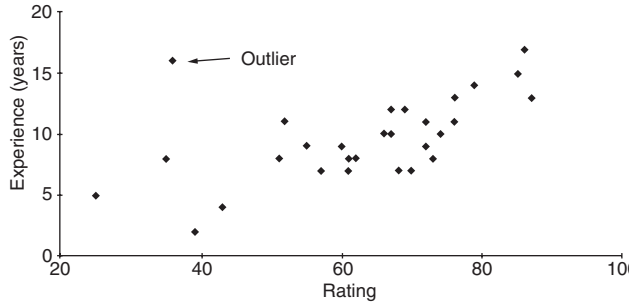


FIGURE 9.7 Outliers outside pattern of explanatory variables.

CHECKING ASSUMPTIONS WITH RESIDUALS

Residuals are the primary tools for checking whether the assumptions necessary for linear regression appear to be violated. We can draw histograms of the residuals, plot them against the observed and predicted values, recompute them excluding certain cases, and manipulate them in other ways. By examining the resulting plots and statistics, we can learn much about how appropriate the regression model is for a particular data set. In the next few sections, we will consider how to check each of the assumptions in turn.

Normality

If the relationship is linear and the dependent variable is normally distributed for each value of the independent variable (in the population), then the distribution of the residuals should also be approximately normal. A simple histogram can demonstrate this.

On the other hand, when the distribution of residuals does not appear to be normal, you can sometimes transform the data to make it appear more normal. When you "transform" a variable, you change its values by taking square roots, or logarithms, or some other mathematical function of the data. If the distribution of residuals is not symmetric but has a tail in the positive direction, it is sometimes helpful to take logs of the dependent variable. If the tail is in the negative direction and all data values are positive, taking the square root of the data may be helpful.

The distribution of your residuals may appear not to be normal for several reasons besides a population in which the distributions are not normal. If you have a variance that is not constant for different values of the independent variable, or if you simply have a small number of residuals, your histogram may also appear not to be normal. So it is possible that after you have remedied some of these problems, the distribution of residuals may look more normal. To check whether the variance appears to be constant, you can plot the residuals against the predicted values and also against the values of the independent variable.

You may find some common transformations useful when the variance does not appear to be constant. If the variance increases linearly with the values of the independent variable and all values of the dependent variable are positive, take the

square root of the dependent variable. If the standard deviation increases linearly with values of the independent variable, try taking logs of the data.

Linearity

To see whether it is appropriate to assume a linear relationship, you should always plot the dependent variable against the independent variable. If the points do not seem to cluster around a straight line, you should not fit a linear regression model. Another way to see whether a relationship is linear is to look at the plots of the residuals against the predicted values and the residuals against the values of the independent variable. If you see any type of pattern to the residuals — that is, if they do not fall in a horizontal band — you have reason to suspect that the relationship is not linear.

Sometimes when the relationship between two variables does not appear to be linear, it is possible to transform the variables and make it linear. Then you can study the relationship between the transformed variables using linear regression.

It may seem that when you transform the data, you are cheating or at least distorting the picture. But this is not the case. All that transforming a variable does is change the scale on which it is measured. Instead of saying that a linear relationship exists between work experience and salary, you say that a linear relationship exists between work experience and the log of salary. It is much easier to build models for relationships that are linear than those that are not. That is why transforming variables is often a convenient tactic.

How do you decide what transformation to use? Sometimes you might know what the mathematical formula is that relates two variables. In that case, you can use mathematics to figure out what transformation you need. This situation happens more often in engineering or the physical or biological sciences than in the social sciences. If the true model is not known, you choose a transformation by looking at the plot of the data. Often, a relationship appears to be nearly linear for part of the data but is curved for the rest. The log transformation is useful for "straightening out" such a relationship. Sometimes taking the square root of the dependent variable may also straighten a curved relationship. These are two of the most common transformations, but others can be used.

When you try to make a relationship linear, you can transform the independent variable, the dependent variable, or both. If you transform only the independent variable you are not changing the distribution of the dependent variable. If it was normally distributed with a constant variance for each value of the independent variable, that remains unchanged. However, if you transform the dependent variable, you change its distribution. For example, if you take logs of the dependent variable, then the log of the dependent variable — not the original dependent variable — must be normally distributed with a constant variance. In other words, the regression assumptions must hold for the variables you actually use in the regression equation.

Independence

Another assumption that we made was that all observations are independent. (The same person is not included in the data twice on separate occasions. One person's

values do not influence the others'.) When data are collected in sequence, it is possible to check this assumption. You should plot the residuals against the sequence variable. If you see any kind of pattern, you should be concerned.

Finally, it is important to examine the data for violation of the assumptions since significance levels, confidence intervals, and other regression tests are sensitive to certain types of violations and cannot be interpreted in the usual fashion if serious departures exist. If you carefully examine the residuals, you will have an idea of what sorts of problems might exist in your data. Transformations provide you with an opportunity to try to remedy some of the problems. You can then be more confident that the regression model is appropriate for your data.

How can you tell whether the assumptions necessary for a regression analysis appear to be violated? Here are the key points to remember:

- A residual is the difference between the observed value of the dependent variable and the value predicted by the regression model.
- To check the assumption of normality, make a histogram of the residuals. It should look approximately normal.
- To check the assumption of constant variance, plot the residuals against the predicted values and against the values of the independent variable. There should be no relationship between the residuals and either of these two variables. If you note a pattern in the plots, you have reason to suspect that the assumption of constant variance is violated.
- To check whether the relationship between the two variables is linear, plot the two variables. If the points do not cluster about a straight line, you have reason to believe that the relationship is not linear.
- If any of the assumptions appear to be violated, transforming the data may help. The choice of the transformation depends on which assumption is violated and in what way.

MULTIPLE LINEAR REGRESSION

A special class of statistical techniques, called *multivariate methods*, is used for studying the relationships among several interrelated variables. The goals of such multivariate analyses may be quite different from that of a univariate analysis, but they share many common features. Here we will take a look at some of the most popular ones.

You can use *multiple linear regression* analysis to study the relationship between a single dependent variable and several independent variables of the form.

$$Y = \text{Constant} + B_1X_1 + B_2X_2 + B_3X_3 \ldots B_nX_n$$

The model looks like the regression model we already have seen. The difference is that you now have several variables on the independent variable side of the model. The independent variables are indicated by X_1, X_2, X_3, …, X_n and the coefficients by B_1, B_2, B_3, …, B_n. As before, the method of least squares can be used to estimate all of the coefficients.

Perhaps the most important issue with multiple linear regression is that because the variables are measured in different units, you cannot just compare the magnitudes of the coefficients to one another. Because of this characteristic, the experimenter must standardize the variables in some fashion. That standardization takes place with a statistic called BETA and contains the regression coefficients when all variables are standardized to a mean of zero and a standard deviation of one. Of course, just like before we still use the significance test for the test of the null hypothesis that the value of a coefficient is zero in the population. You can see that the null hypothesis can be rejected for all of the variables.

When you build a model with several interrelated independent variables, it is not easy to determine how much each variable contributes to the model. You cannot just look at the coefficients and say *this* is an important variable for predicting the dependent variable and *this* one is not. The contributions of the variables are "shared." The goodness-of-fit statistics we considered for a regression model with one independent variable can easily be extended to a model with multiple independent variables.

SELECTING INDEPENDENT VARIABLES

Often you do not know which independent variables together are good predictors of the dependent variable. You want to eliminate variables that are of little use from your equation so you will have a simple, easy-to-interpret model. You can do this with the assistance of what are called *variable selection methods*. Based on statistical considerations, such as the percent of variance explained by one variable that is not explained by any other variables, some software packages (for example SPSS) select a set of variables for inclusion in a regression model. Although such procedures often result in a useful model, the selected model is not necessarily best in any absolute sense.

DISCRIMINANT ANALYSIS

You can use regression analysis to predict values of the dependent variable based on a set of independent variables. The dependent and independent variables are all measured on an interval or ratio scale. What if the dependent variable is not measured on an interval or ratio scale? Instead, your data for the dependent variable are ordinal. What you might want to use in this case is a procedure called *discriminant analysis*.

In discriminant analysis, you compute "discriminant scores" for each case to predict what group it is in. These scores are obtained by finding linear combinations of the independent variables. (A linear combination is formed by multiplying each variable by some constant and then adding up the products.) For example, you might compute an individual's score by taking 2 times income in thousands, plus 1.5 times age, plus 1.1 times education. Discriminant analysis uses mathematical techniques to determine the way of computing scores that results in the best separation among the groups (in other words, the most accurate prediction of what group each case is in). Statistical packages such as SPSS do all the computations for you, including selection of the best coefficients. The result of a discriminant analysis tells you how

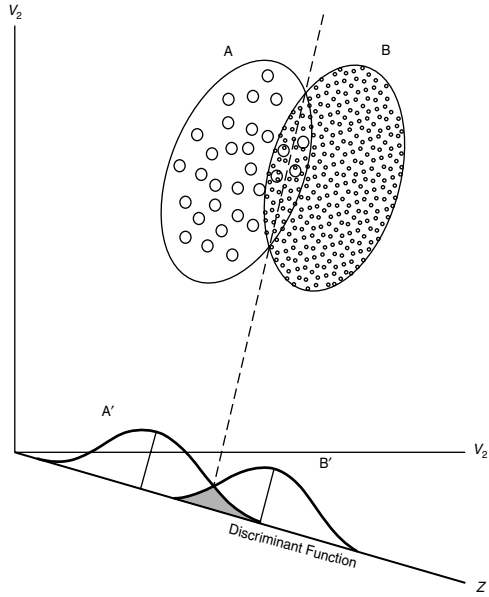


FIGURE 9.8 Graphical illustration of two-group discriminant analysis.

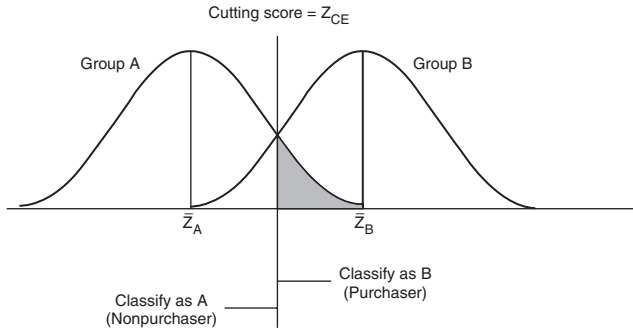


FIGURE 9.9 Optimal cutting score with equal sample sizes.

well you are able to predict what case a group falls into, based on the values of the independent variables.

Graphical representations of discriminant analysis are shown in Figures 9.8 through 9.12.

LOG-LINEAR MODELS

Using a cross-classification table and the chi-square statistic, you were able to test whether two variables that have a small number of distinct values are independent. However, what if you wanted to know the effect of additional variables on the relationships that you are examining? You could always make a cross-tabulation

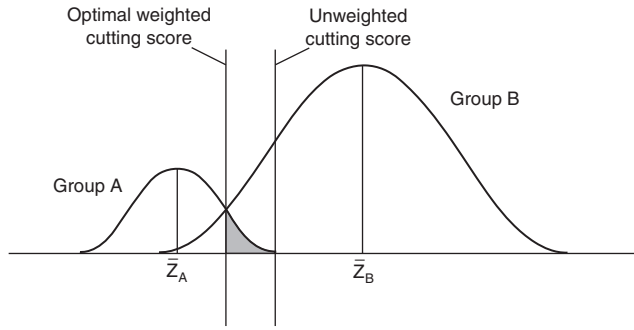


FIGURE 9.10 Optimal cutting score with unequal sample sizes.

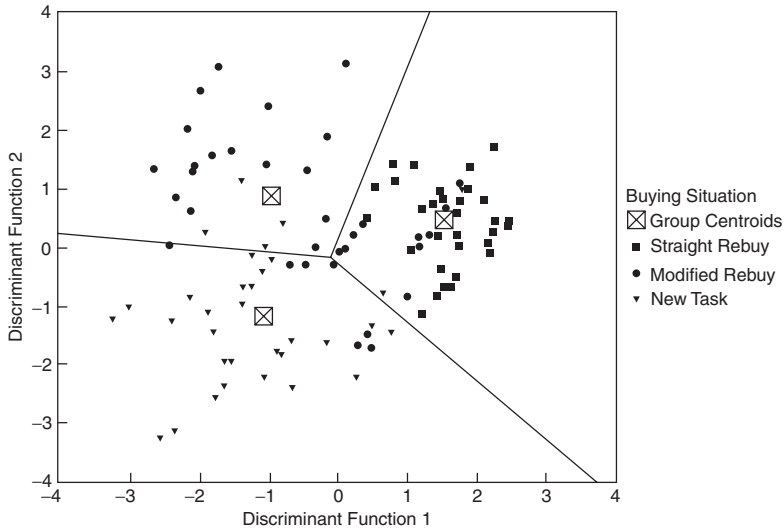


FIGURE 9.11 Territorial map and rotated discriminant Z scores.

table of all of the variables, but this would be very difficult to interpret. You would have hundreds or thousands of cells and most of them would contain few cases, if any.

One way to study the relationships among a set of categorical variables is with log-linear models. With a log-linear model you try to predict the number of cases in a cell of a cross-tabulation, based on the values of the individual variables and on their combinations. You see whether certain combinations of values are more likely or less likely to occur than others. This tells you about the relationships among the variables.

FACTOR ANALYSIS

If you give a group of students 100 different aptitude tests, their scores on the different tests will no doubt be correlated. The tests probably measure some of the

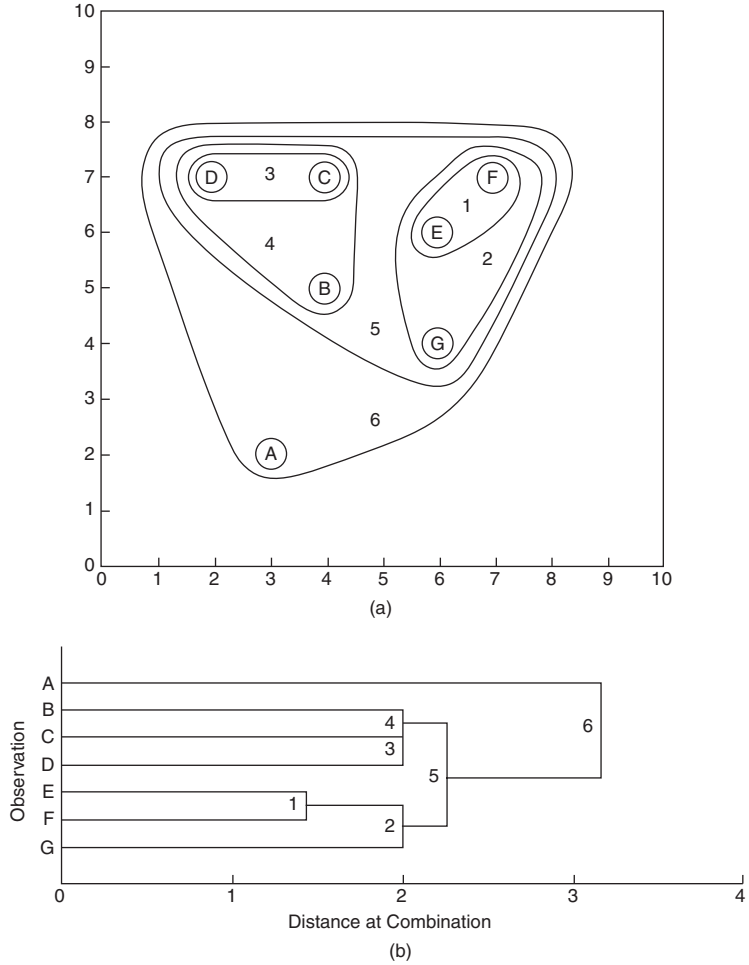


FIGURE 9.12 Graphical portrayals of the hierarchical clustering process: (a) nested groupings, (b) dendogram.

same characteristics, such as verbal skills, mathematical aptitude, reasoning ability, and perceptual speed. Characteristics such as "verbal skill," "mathematical aptitude," and "reasoning ability" are not well-defined, easily measurable variables like weight or age. Instead they can be thought of as unifying concepts or labels that characterize responses to related groups of variables. A mathematically apt person would score well on all of the tests related to mathematical skills. In fact, that is how the definition of mathematical aptitude was formulated. *Factor analysis* is a statistical technique that attempts to measure such concepts.

In some research situations, you have a set of interrelated variables such as consumer ratings of products. You think that the ratings are correlated because people are rating the products on similar dimensions such as product quality and utility. But you do not know what these underlying dimensions, or factors, are. You can

use factor analysis to help you identify these underlying concepts by using a number of variables that you can directly measure.

CLUSTER ANALYSIS

If you question sick patients about their symptoms, you will undoubtedly have a very long list of complaints. You may think that you will find as many combinations of symptoms as you have patients. However, if you study the types of symptoms that frequently occur together, you will probably be able to put the patients into groups — those who have respiratory disturbances, those who have gastric problems, those who have cardiac difficulties. Classifying the patients into groups of similar individuals may be helpful both for determining treatment strategies and for understanding how the body malfunctions.

In statistics, the search for similar groups of objects or people is called *cluster analysis*. By forming clusters of objects and then studying the characteristics the objects share, as well as those in which they differ, you can gain useful insights. For example, cluster analysis has been used to cluster skulls from various archeological digs into the civilizations from which they originated. Cluster analysis is also frequently used in market research to identify groups of people for whom various marketing pitches may be particularly attractive.

TESTING HYPOTHESES ABOUT MANY MEANS

Previously, we tested the hypothesis about the equality of population means. We wanted to know if people find the ride comfortable, normal, or rough. We used the analysis of variance procedure to test hypotheses that more than two population means are equal. We tested whether there was a difference in education among the three excitement groups.

What if we have several interrelated dependent variables, such as education and income, about which we wish to test hypotheses? How can we test hypotheses that *both* education and income do not differ among the three excitement groups in the population? *Multivariate analysis of variance*, or MANOVA, is used to test such hypotheses. Using MANOVA, you can compare four instructional methods based on student achievement levels, satisfaction, anxiety, and long-term retention of the material. Or you compare five new ice cream flavors based on the amount consumed, a preference rating, and the price people say they would pay.

If the same variable is measured on several different occasions, special "repeated measures" analysis of variance techniques can be used to test hypotheses. These can be thought of as extensions of the simple paired t test.

The previous chapters explained some of the more widely used statistical techniques, and this chapter has attempted to give an idea of the more sophisticated methods available. Still others exist. You can use nonparametric procedures that do not require such stringent assumptions about the distributions of variables. You can use procedures for analyzing specialized types of data such as test scores or survival times. There are often many different ways to look at the same problem. No one way is best for every problem; each view tells you something new.

SELECTED BIBLIOGRAPHY

Brown, H. and Prescott, R., *Applied Mixed Models in Medicine*, John Wiley & Sons, New York, 2000.

Chatterjee, S., Hadi, A.S., and Price, B., *Regression Analysis by Example*, John Wiley & Sons, New York, 1999.

Cook, R.D., *Regression Graphics: Ideas for Studying Regressions through Graphics*, John Wiley & Sons, New York, 1998.

Cook, R.D. and Weisberg, S., *Applied Regression Including Computing and Graphics*, John Wiley & Sons, New York, 1999.

Draper, N.R. and Smith, H., *Applied Regression Analysis*, 3rd ed., John Wiley & Sons, New York, 1998.

Freund, R.J. and Littell, R.C., *SAS System for Regression*, John Wiley & Sons, New York, 2000.

Hosmer, D.W. and Lemeshow, S., *Applied Logistic Regression*, 2nd ed., John Wiley & Sons, New York, 2000.

Khattree, R. and Naik, D.N., *Applied Multivariate Statistics with SAS Software*, John Wiley & Sons, New York, 1999.

Khattree, R., Naik, D.N., and SAS Institute Inc., *Multivariate Data Reduction and Discrimination with SAS Software,* John Wiley & Sons, New York, 2000.

Khuri, A., Mathew, T., and Sinha, B.K., *Statistical Tests for Mixed Linear Models*, John Wiley & Sons, New York, 1998.

McCulloch, C.E. and Searle, S.R., *Generalized, Linear, and Mixed Models*, John Wiley & Sons, New York, 2000.

McLachlan, G. and Peel, D., *Finite Mixture Models*, John Wiley & Sons, New York, 2000.

Rencher, A.C., *Linear Models in Statistics*, John Wiley & Sons, New York, 1999.

Ryan, T.P., *Modern Regression Methods*, John Wiley & Sons, New York, 1996.

Seber, G.A.F. and Wild, C.J., *Nonlinear Regression*, John Wiley & Sons, New York, 1989.

Schimek, M.G., *Smoothing and Regression: Approaches, Computation and Application*, John Wiley & Sons, New York, 2000.

10 Common Miscellaneous Statistical Tests

This chapter will discuss common tests for nominal data (binomial, chi square [I], chi square [II], McNemar, and the Cochran Q), ordinal data (Kolmogorov-Smirnov, Mann-Whitney U, sign, Wilcoxon, Kruskal-Wallis, and Friedman), and interval data (*t* test [I], *t* test [II], *t* test [III], and Scheffe's test).

BINOMIAL TEST

Some of the statistical tests you will be studying are used to analyze data with many categories or outcomes. However, when only two categories are present, the binomial test (sometimes called a test of proportion) is applicable. The two categories could be, for example: grades in a pass/fail course, party affiliation as broken down into either Republican or Democrat, evaluation of a product as good or bad, a decision about a process as go or no go, outcome in tossing a coin heads or tails, or outcome in rolling a six or not a six on a die.

The proportion of cases in one category is referred to as P and in the other category as Q. The value of P + Q always equals one. If you know the value of P, you find Q by subtracting P from one. The requirements for the binomial test are:

1. Nominal data
2. One-group test
3. Two categories only
4. Sample size can be less than five
5. Independent observations
6. Simple random sample
7. Data in frequency form

The general formula for the binomial is:

$$P(x) = \binom{N}{x} P^x Q^{N-x}$$

The terms $\binom{N}{x}$ are binomial coefficients and are computed by the following formula:

$$\frac{N!}{x!(N-x)!}$$

The meanings of the symbols are as follows: ! = factorial, N = number of trials or sample size, X = number of favorable outcomes for a series of trials, P = probability of favorable outcome in a single trial, Q = (1 – P) = probability of unfavorable outcome in a single trial.

REMARKS ON THE BINOMIAL TEST

The binomial is a useful test because it computes exact probabilities in order to get the region of rejection, and you can use either a one- or two-tailed test. If the number in the sample is greater than 25, however, calculations of this type become very cumbersome. The chi-square test and others like it do not compute exact probabilities, but they are easier to calculate.

CHI-SQUARE (I) TEST

The chi-square (I) test (also known as "goodness-of-fit" test) is used to determine whether a significant difference exists between the expected frequencies and the observed frequencies in one or more categories. Do the number of individuals or objects that fall in each category differ significantly from the number you would expect? Is this difference between the expected and observed due to sampling error, or is it a real difference? The requirements for the chi-square (I) test are:

1. Nominal data
2. One-group test
3. One or more categories
4. Independent observations
5. Adequate sample size
   a. The expected frequencies should be sufficiently large for two categories five or larger.
   b. When more than two categories are present, no more than 20% should be smaller than five.
6. Simple random sample
7. Data in frequency form
8. All observations must be used
9. Two-tailed test only (This test cannot be used as a one-tailed test. If directional testing is necessary, you should use a different test.)

The chi-square formula is:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O = observed frequencies in each category, E = expected frequencies for the corresponding observed frequencies, Σ = sum of, k = number of categories, and df = degrees of freedom (number of categories minus one [k – 1]).

CHI-SQUARE (II) TEST

The last section dealt with chi-square and its use with one group. However, chi-square has wide application, and it can be used with two or more groups. To distinguish between the use of chi-square with one group and with two or more groups, we shall use the terms chi-square (I) and chi-square (II). The question we ask in discussing chi-square (II) is: do two or more groups differ in respect to some characteristics? In other words, do the number of frequencies that fall into each category for one group differ significantly from the number that fall into each category for another group or groups? The requirements for chi-square (II) are:

1. Nominal data
2. Two or more groups
3. Independent observations
4. Adequate sample size
   a. No more than 20% of expected frequencies can be smaller than five.
   b. When expected frequencies are very small, use the Fisher exact probability test (Siegel, 1956), not the chi-square (II).
5. Two-tail test only

The method for finding chi-square (II) differs from that for chi-square (I) in the way the expected frequencies are found. You have no prior basis for computing the expected frequencies, so they are derived from the data. This involves setting up a contingency table. The formula however, is the same as that for (I).

A WORD OF CAUTION ON χ^2

Chi-square is a convenient measure of association between two factors when the factors are not quantitative. It indicates the degree to which the frequencies in a cross-tabulation of the two factors deviate from what they would be if no interrelation existed between the factors. The computed chi-square has a specific level of statistical significance that you can look up in a standard table.

Suppose we ask 300 testers to rate two brands of a product (A and B) both in terms of overall preference and preference regarding "comfort." By a convenient coincidence, the "comfort" preference divides exactly even, with 100 preferring A, 100 preferring B, and 100 having no preference.

Comfort level	A	B	No preference
A	70	30	100
No preference	55	45	100
B	40	60	100
Total	165	135	300

Clearly, a strong association exists between the preference on "comfort" and overall preference; chi-square is 18.4, indicating a significance level of 99%+. However, let us assume that we suspect the results and upon further investigation, we find that we have recorded the data in the wrong cell. The table should have looked like:

Comfort level	A	B	No preference
A	70	30	100
B	55	45	100
No preference	40	60	100
Total	165	135	300

Now, let us see what we have. It still looks like a strong association for A but not for B, so we should have a lower chi-square, right?

No. Chi-square is still 18.4. As long as the numbers stay the same, it does not matter how they are labeled. Like the scarecrow in *The Wizard of Oz,* chi-square does not have a brain. It is merely an algorithm, a mechanical process based on numbers regardless of what they represent. By itself, it never can take the place of a regression or correlation because it cannot describe the relationship; it can only gauge its *statistical* significance, entirely regardless of logic or sense.

Chi-square is nonparametric. To describe a relationship in numerical terms, we need numerical values — that is, parameters. If we arbitrarily assign value +1 to preference for A and –1 to preference for B, we can compute a correlation coefficient of r = +.246 for the original tabulation and exactly half that for the corrected distribution. The parametric regression/correlation, unlike chi-square, is affected by the way the rows and columns are labeled because each label has a specific value.

So chi-square is a very useful index when we cannot assign values, but it is very easy to misuse it; it does not have a brain, so the analyst has to use his or her own brain to interpret it correctly.

McNEMAR TEST

The McNemar test is used for before-and-after research designs. It is used with matched pairs or when a subject is his/her own control. The purpose is to determine the significance of the observed change. The requirements are listed below.

1. Nominal data
2. Two groups

3. Related groups
4. Expected frequencies five or more
5. Two-tailed test only (The chi-square table is used for the critical value.)

The formula for the McNemar test is:

$$\chi^2 = \frac{(|A - D - 1|)^2}{A + D}$$

where A = number of individuals changing in one direction, D = number of individuals changing in the opposite direction from A, and E = expected number of individuals under the null hypothesis ½ (A + D). Df = degrees of freedom is always 1 for the McNemar test.

COCHRAN Q TEST

The McNemar test for two related groups can be extended to more than two groups. This extension is called the Cochran Q test. It is a method for testing whether three or more matched sets of frequencies differ significantly. The matching can be on relevant features of different people, or the same person can be used under different conditions.

Scores for the Cochran Q test can take only two values: zero or one. The value one represents values that are recorded as positive, while the zero represents those that are negative. The requirements for the Cochran Q test are as follows:

1. Nominal data
2. Three or more groups
3 Related groups
4. Data in frequency form
5. Two-tail test only

The Cochran Q formula is:

$$Q = \frac{(k-1)\left(k\sum C^2 - T^2\right)}{KT - \sum R^2}$$

where k = number of groups, R = row totals, C = column totals, and T = grand total (sum of R or sum of C).

KOLMOGOROV-SMIRNOV TEST

The Kolmogorov-Smirnov is an ordinal test used with one-group samples. The experimenter uses it to find out whether a distribution of observations is significantly

different from a theoretical distribution. The test compares the cumulative distribution of the observed scores and the cumulative distribution of the expected scores. The point where the two distributions show the largest divergence is then determined. Next, the experimenter refers to the sampling distribution to determine whether this divergence is the result of chance or a real difference. Do the scores in these two distributions come from the same population? To test this difference, one uses the critical value of D for the Kolmogorov-Smirnov test. The requirements for the Kolmogorov-Smirnov test are as follows:

1. Ordinal data
2. One group
3. Simple random sample

The formula for the Kolmogorov-Smirnov test is:

$$D = \frac{LD}{N}$$

where LD = large difference, N = number of individuals in the sample, O = number of individuals observed, E = number of individuals you would expect in the sample, OC = observed cumulative distribution, EC = expected cumulative distribution, and f = frequency of scores.

Special note*:* In the process of calculating this test, notice that the definitions of OC and EC are part of the *cumulative distribution.* Before you can use the Kolmogorov-Smirnov test, you must know how to create a cumulative distribution.

USE OF THE MANN-WHITNEY U TEST

The Mann-Whitney U can be applied when you have two independent, randomly selected groups of unequal sizes. It tests whether two independently drawn samples have been drawn from the same population. To do the comparison, the experimenter must rank the data. Once that is done, then the question is: Do the scores from one sample significantly differ from those of the other sample, so that we can conclude each sample represents a different population, or is the difference due to the luck of the draw? In the latter case we would conclude that, though the samples may differ, they represent the same population. The decision is made on the critical value as identified in a special table of values for this test. The requirements for the Mann-Whitney U are as follows:

1. Ordinal data
2. Two groups
3. Independently drawn samples
4. Data in ranks
5. Simple random samples
6. Sample size can be different for the two groups.

When the n_2 is between 9 and 20 then the formulas for the test are:

1. Mann-Whitney U

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$

2. Mann-Whitney U

$$U_2 = n_1 n_2 - U_1$$

The reason that two formulas are necessary is that we need to find the smaller U of the two. Since we are dealing with two populations, usually of different sizes, it stands to reason the two formulas will give different values of U. We perform both tests for U, and whichever happens to be smaller is used as our U value. Do not worry if you do not understand these formulas now; just look at them. The symbols we use for computing and understanding the formulas are n_1 = size of smaller group, n_2 = size of larger group, N = total number in both groups ($n_1 + n_2$), ΣR_1 = sum of the ranks in group one, ΣR_2 = sum of the ranks in group two.

When n_2 is larger than 20, then the Mann-Whitney U test is:

$$Z = \frac{U + \frac{1}{2} - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

Obviously, when n_2 is larger than 20, you have to carry out some further calculations. In this case you use your value of U to compute the value of z given by the formula. The sampling distribution for z is approaching the normal distribution, and as a consequence, the table values for the z are appropriate. That means that you reject the null hypothesis if the probability for z is equal to or less than the predetermined level of significance. The one-tailed probabilities are given in the z table; for two-tailed probabilities double the table values.

COMMENT ABOUT THE MANN-WHITNEY U

The Mann-Whitney U test, a nonparametric test, is often used as a substitute for the *t* test, a parametric test. This test is used when the *t* test's assumptions appear to be in doubt — for example, when the shape of the distribution does not appear to be normal and the level of measurement is ordinal. The Mann-Whitney U is an easy-to-apply nonparametric test.

SIGN TEST

Earlier we discussed the binomial test. You used the binomial distribution to test different hypotheses. Now we are going to study the sign test, which is based on

the binomial distribution. When the sample size in our experiment is 25 or smaller, we can evaluate the results by using the binomial distribution as the sampling distribution.

The sign test measures the significance of the difference between two treatment conditions. It is used with two groups that are matched. You choose this test when you are concerned with the direction of differences, which you indicate by a plus or a minus sign. When using the sign test you analyze every matched pair's scores, recording whether the sign of the difference between a pair is negative or positive. If the scores of a pair are the same, you disregard that pair. The pertinent requirement for this test is that each pair is matched on the important variable that you are studying. The actual test of significance is based on the chi-square table (if fewer than 25 samples) and the z statistic if the sample is larger than 25. The requirements for the sign test are as follows:

1. Ordinal data
2. Two-group test
3. Related groups
4. When a pair of observations are tied, neither is used
5. Plus and minus signs are used to indicate differences

What we are trying to find out is whether the number of plus signs exceeds the number of minus signs. If the two are the same, then there is no difference. The symbols used are: x = the number of pluses or the number of minuses, whichever is less, and N = total number of pluses and minuses in the group.

The frequency of occurrence of a particular plus or minus sign can be determined by looking at the probabilities table for the binomial distribution, when p and q both are equal to $^1/_2$.

COMMENTS ABOUT THE SIGN TEST

The major weakness of the sign test is that it does not use all the information given, so if possible a parametric test should be used. As already mentioned, when the sample size is larger than 25, the normal distribution approximates the binomial sampling distribution, and the formula used is:

$$Z = \frac{(x \pm .5) - \frac{1}{2}N}{\frac{1}{2}\sqrt{N}}$$

The table for the z values is used for interpretation.

WILCOXON SIGNED-RANKS TEST

The Wilcoxon test compares the distribution groups. The groups are related in one of two ways:

1. The same objects/people have been tested under two different conditions.
2. Pairs of objects/people have been matched on the same basis before being tested.

The purpose of this test is to determine whether the differences between the groups favor one group over the other. The sign test, which you have just read about, concerns itself with only the direction of the differences between the pairs; it does not consider the size of the differences. All we recorded by using a plus or a minus sign for each matched pair of scores for the sign test was whether one of the pairs was larger or smaller; we did not record how large or small this difference was. The sign test does not use all the information that is available. However, the Wilcoxon test concerns itself not only with the direction of the differences but also with their size. It makes a distinction between these differences by ranking them. The critical value for this test can be found on a table for this particular test. The requirements of the Wilcoxon signed-ranks test are as follows:

1. Ordinal data
2. Two groups
3. Related groups
4. Ranked data

The symbols used for this test are: N = number of matched pairs, excluding those with a deviation (D) of zero; T = the smaller value for either ΣR^+ or ΣR^-.

SAMPLE SIZES LARGER THAN 25

When you have a sample size larger than 25, you cannot use the Wilcoxon table. In such instances, your T is almost normally distributed. When this is the case, you proceed as before to find T. That is, you find the sum of the smaller ranks. After you find T, you use the following formula, which you evaluate based on the z table.

$$Z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}$$

KRUSKAL-WALLIS TEST

The Kruskal-Wallis test is similar to the Mann-Whitney U test. The Mann-Whitney U compared two groups, whereas the Kruskal-Wallis compares three or more groups. The experimenter wants to know if the difference among the groups is due to sampling error or is a real difference. The Kruskal-Wallis test determines whether a difference is present by finding out if the sums of the ranks for each of its groups differ significantly from each other. A significant value of H (based on the chi-square table or the Kruskal-Wallis table) implies that the medians of the distribution are

not the same. When using the Kruskal-Wallis test, the experimenter does not have to be concerned about whether the test is one- or two-tailed. The only concern is whether a difference exists. The requirements for using this test are:

1. Ordinal level
2. Three or more groups
3. Independent groups
4. Simple random sample

The formula for the calculation is:

$$H = \frac{12}{N(N+1)}\left[\frac{\left(\sum R_1\right)^2}{n_1} + \frac{\left(\sum R_2\right)^2}{n_2} + \frac{\left(\sum R_3\right)^2}{n_3} + \ldots + \frac{\left(\sum R_k\right)^2}{n_k}\right] - 3(N+1)$$

The symbols used in calculating the test are: N = total number of objects ($n_1 + n_2 + n_3 + \ldots$), Df = k – 1 Degrees of freedom, equal to the number of groups minus one, n_1, n_2, n_3, … = number of objects in group 1, group 2, group 3 and so on, $(\Sigma R_1)^2$, $(\Sigma R_2)^2$, $(\Sigma R_3)^2$, … = sum of the ranks for group 1 squared, group 2 squared, group 3 squared, and so on.

THE EFFECT OF TIES

When the number of ties is large, H will be a little smaller than it should be. There is a correction for ties that makes it easier to reject the null hypothesis. If you do not correct for ties, your rejection of the null hypothesis will be harder, and your test is more conservative. We shall not deal with this correction formula because it has little effect on the value of H unless the number of ties is extreme. If you are interested in the correction, consult Hays (1973) and Siegel (1956).

FRIEDMAN TEST

The Friedman test is useful when we want to test the null hypothesis that many groups have been drawn from the same population. The Friedman is an extension of the previously discussed Wilcoxon test. The Wilcoxon can be used with only two groups, while the Friedman can be used with three or more groups. Like the Wilcoxon, the Friedman uses related or matched groups. Either these groups are matched on the basis of some variable, or the same subjects are used for different treatments. Since the Friedman uses groups that are matched, the number of subjects in each sample is the same. Significance is measured based on the chi-square table. The requirements for this test are:

1. Ordinal data
2. Three or more groups

3. Related groups
4. Sample drawn at random from matched scores

The formula for this statistic is:

$$\chi_r^2 = \frac{12}{Nk(k+1)}\left[\sum R_i^2\right] - 3N(k+1)$$

The symbols used in the calculating process of the statistic are: k = number of columns, N = number of rows, ΣR_i^2 = sum of the squared rank sums, df = k – 1 degrees of freedom, equal to the number of columns minus one.

t TEST

The *t* test, often referred to as Student's *t* or the *t* ratio, was first described in 1908 in an article in *Biometrika* by William S. Gosset, who wrote under the pseudonym "Student." Gosset, a 32-year-old chemist, was a consultant for the Guinness Brewery in Ireland. (Here we have taken some liberties. The *t* ratio will be known as *t* [I], *t* [II], and *t* [III]. When the *t* is used with one group it is referred to as *t* [I]; with two independent groups, as *t* [II]; and with two related groups, as *t* [III].) In doing his research he discovered the *t* distribution, which permitted him to test hypotheses with small samples when the population had a normal distribution, the population mean was assumed, and the population standard deviation was unknown.

In the majority of small practical research problems, knowledge of the standard deviation is not available. If previous experience does not supply the standard deviation, one uses the *t*. In using the *t* to test a hypothesis, you calculate the *t* and then refer to a *t* table to determine the probability for this statistic.

t Test (I): The *t* test (I) is used with one-group samples where the population has a normal distribution. For example, suppose a population is assumed to have a mean of 1.85 meters and it has a normal distribution. We then reach into this population, pull out a sample, compute its mean, and compare its mean to the population mean of 1.85 meters by using the *t* test (I). If the difference between the sample mean and the assumed population mean is too large, then the assumption that the population mean is 1.85 meters is rejected. The requirements for the *t* test (I) are as follows:

1. At least the interval level of measurement
2. Random sample
3. Sample drawn from a population that has a normal distribution

The formula for the *t* test (I) is

$$\frac{(\overline{X} - \mu)}{s_{\overline{X}}}$$

that is, the sample means minus the population mean divided by the standard error of the mean. Using the formula for *t* does not require skills beyond what we already have learned. In effect, you are calculating a mean, a standard deviation, and a standard error of the mean. By using these three formulas, you can compute a *t* for any group of scores. To shorten this task, a formula that combines all the components is used:

$$t = \frac{\overline{X} - \mu}{\sqrt{\frac{\sum X^2 - \frac{\left(\sum X\right)^2}{N}}{N(N-1)}}}$$

The symbols you will encounter in learning about *t* (I) are $\overline{X}$ = sample mean, μ = population mean, ΣX^2 = square each score, then find the sum of the squares, N = count the number of scores, $(\Sigma X)^2$ = square the sum of the scores, and df = N – I = degrees of freedom, number of scores minus one.

COMMENTS ABOUT THE *T* DISTRIBUTION

The distribution for the *t* is very similar to the normal curve; that is, it is symmetrical with a mean of zero and a standard deviation slightly more than one. However, the *t* curve is more peaked than the normal curve. Furthermore, the *t* is not just one distribution, like the normal curve, but many distributions. Each of these distributions looks different, depending upon its degrees of freedom. Figure 10.1 shows three examples of the *t* curve with different degrees of freedom. Figure 10.2 shows the *t* and standard normal distributions.

Notice that the more degrees of freedom you have, the more your *t* resembles the normal curve. When the number in your sample is equal to or greater than 100, the normal curve may be substituted for the *t* distribution. Because the *t* is composed of so many distributions, tables for the *t* include these distributions. However, most

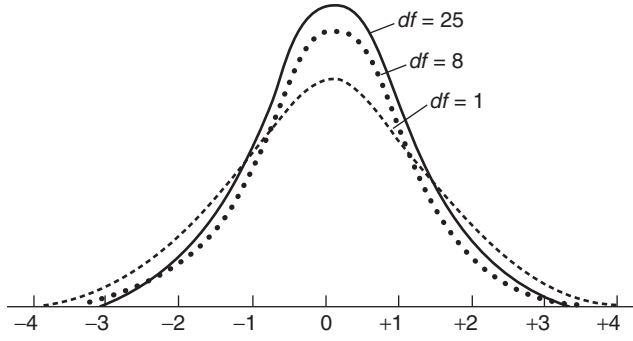


FIGURE 10.1 *t* Distributions with 1, 8, and 25 df.

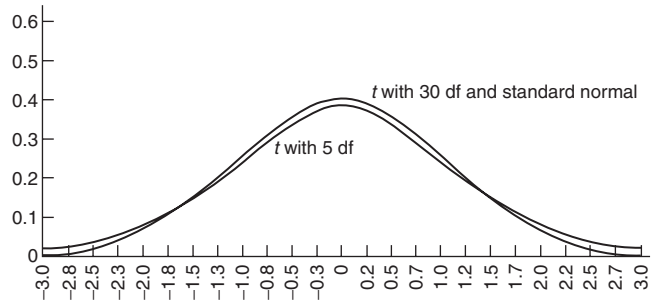


FIGURE 10.2 The *t* and standard normal distributions.

tables are abbreviated, because you would need a volume to show the different distributions for the *t*.

T TEST (II)

The *t* test (II) concerns itself with two groups. These groups are independent; that is, an individual in one group cannot be in any way related to an individual in the other group. The purpose of the *t* (II), or *t* ratio as it is called, is to determine whether the mean of one group is significantly different from the mean of the other group. The requirements of the *t* (II) are as follows:

1. The two groups are independent.
2. Measurement is at least at the interval level.
3. The populations are both normally distributed.
4. The populations have the same variances. (Older statistics books and articles suggest that a test of homogeneity of variance be carried out before it is permissible to use a *t* ratio. However, many modern statisticians feel that such a test is not worth the time and effort. Where these tests for homogeneity of variance are most needed — on small samples with unequal sizes — they are the least effective [Hays, 1973].)
5. The samples are drawn at random.

IMPORTANCE OF REQUIREMENTS THREE AND FOUR

Requirement Three states that the two populations should be normally distributed. A severe departure from normality seems to have little effect on the conclusions when sample sizes are 30 or more. Of course, the results would be more accurate if the distribution were normal. In reality you can violate this assumption and not worry about it as long as your samples are not extremely small (Hays, 1973).

Requirement Four (homogeneity of variance) states that the populations should have the same variances. When the populations do not have the same variances and the sample sizes are equal, there is little effect on the conclusions reached by the *t* test. However, when the sample sizes are extremely small and of unequal sizes, the

t ratio is affected. One way to handle this problem is to avoid using samples of unequal sizes. However, if this is not possible, and if the situation warrants it, use a nonparametric test instead of a *t*. If these two solutions are not feasible, the only other possibility is to use a computational formula that computes the standard error of each sample separately and use a corrected number for your degrees of freedom.

The best formula to use for the *t* test (II) is the following:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\left[\sum X_1^2 - \frac{\left(\sum X_1\right)^2}{N_1}\right] + \left[\sum X_2^2 - \frac{\left(\sum X_2\right)^2}{N_2}\right]}{N_1 + N_2 - 2} \cdot \left(\frac{N_1 + N_2}{N_1 N_2}\right)}}$$

The numerator is the actual difference between the means, whereas the denominator is an estimate of the standard error of the difference between the means. The denominator is an estimate of the variability of the difference between the means of the two samples. When you use this formula, you are dividing the observed differences (numerator) by the variation of differences (denominator) that can be expected due to chance. If no significant difference exists between the groups, the ratio will be equal to zero. The further the ratio deviates from zero, the more likely it is that a real difference exists between the groups. The formula for *t* (II) looks very complicated, but these symbols are in fact old friends. Their meanings are: $\overline{X}$ = mean of group I, ΣX_1^2 = square the individual scores for group I and then find the sum, $(\Sigma X_1)^2$ = sum the individual scores for group I and then square the sum, N_1 = count the number of subjects in group I, and df = $N_1 + N_2 - 2$ degrees of freedom, the number of subjects in group I plus the number of subjects in group II minus 2.

t TEST (III)

The *t* test (III) is also used with two groups. This test's concern is to find out if the mean of one group is actually different from the mean of the other group. The difference between the *t* test (II) and the *t* test (III) has to do with the nature of the two groups. The *t* test (III) is used only in cases where the two groups are related. When we talk about related groups, we mean groups that are matched on some variable or in which the subjects are used more than once. The requirements for the *t* test (III) are as follows:

1. Two groups related
2. At least interval level of measurement
3. Populations both normally distributed
4. Populations having the same variances
5. Samples drawn at random

The formula used to evaluate whether the difference between these two groups is significant is different from the one used for the *t* (II). You compute the differences between each pair of scores and then use this difference to estimate the population standard error of the difference.

$$t = \frac{\overline{D}}{\sqrt{\frac{\sum D^2 - \frac{\left(\sum D\right)^2}{N}}{N(N-1)}}}$$

where $\overline{D}$ = mean of the difference, ΣD^2 = square the differences, then find the sum, $(\Sigma D)^2$ = sum the differences, then square the sum, and N = number of pairs of scores.

SCHEFFE'S TEST

The Scheffe's test is in effect an F test. However, this F test is computed as a separate test for every comparison using the following formula.

$$F = \frac{(\overline{X}_1 - \overline{X}_2)}{MS_w\left(\frac{n_1 + n_2}{n_1 n_2}\right)(k-1)}$$

After this F is computed, you compare it to the table value you found when you did your analysis of variance. If the F is equal to or greater than the table value, it is considered significant. The MS_w is the mean square within from the ANOVA table.

CORRELATION

How do we choose a correlation technique? We choose on the basis of the situation that is being examined and the data's level of measurement. Here we will address three techniques, of which two are nonparametric and one is parametric. These procedures are most often found in beginning statistics books, and each serves a different level of measurement, as illustrated by the following table.

Level of Measurement	Correlational Technique	Kind
Nominal	Contingency coefficient	Nonparametric
Ordinal	Spearman rank	Nonparametric
Interval	Pearson product-moment	Parametric

After we choose a correlation technique, we use it to compute a number called a *coefficient of correlation.* This number tells us the exact strength and direction of

the relationship between the two sets of scores. We begin our discussion with the Pearson product-moment coefficient of correlation.

PEARSON PRODUCT-MOMENT COEFFICIENT

The *Pearson product-moment coefficient of correlation* or the *Pearson r*, as it is sometimes called, was derived by the English statistician Karl Pearson. It is the most popular measure of correlation for measuring the linear relationship between two numerically valued random variables. In order to use this parametric measure, we assume the scores on each variable come from a normally distributed population. The requirements for the Pearson *r* are as follows:

1. Relationship is linear.
2. Scores of the population form a normal distribution curve.
3. Scattergram is homoscedastic.
4. Scores are at interval level of measurement.

WHAT IS THE PEARSON *r*?

The Pearson *r* is simply the mean of the z score products, or

$$r = \sum \frac{z_x z_y}{N}$$

To compute r you do nothing new. You convert every X-score value and every Y-score value to a z score. Then you multiply each z score of X by its z score of Y. You sum the products and divide by the number of pairs. The Pearson *r* then shows you the extent to which individuals have the same position on these two variables. Because you change both sets of scores to z scores, you do not have to worry that the variables are not measured on the same type of scale. In other words, you can correlate weight with height. Looking at this formula and understanding what it means helps one to understand the concept of correlation. However, using this formula to compute *r* is computationally a pain in the neck. Can you imagine the time and effort it would take to convert every X score and Y score to a z score? Needless to say, it makes life easier to know that other formulas have been derived from this basic definitional formula. One of the easiest to use is given below:

$$\text{Pearson r Coefficient} = r = \frac{\left[N\sum XY - \left(\sum X\right)\left(\sum Y\right)\right]}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y^2\right)\right]}}$$

The terms used in this formula have the following meanings: ΣXY = multiply each X by its Y, then sum the results, ΣX = sum of all X, ΣY^2 = square the Y, then

sum the results, ΣX^2 = square the X, then sum the results, ΣY = sum of all Y, and N = number of pairs.

The requirements for using the Pearson *r* are:

1. Relationship is linear.
2. Data of the population form a normal distribution curve.
3. Scattergram is homoscedastic.
4. Data are at interval level of measurement.

The procedure used for the Pearson *r* is:

1. Determine *r* by the formula

$$r = \frac{\left[N\sum XY - \left(\sum X\right)\left(\sum Y\right)\right]}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y^2\right)\right]}}$$

2. Determine the statistical significance of *r* when N is smaller than 30.
   a. Refer to a Critical Value of the *t* value in a table.
   b. Using a df of N – 2, enter table value.
   c. If your *r* is equal to or greater than the table value found in table of *t*, reject the null hypothesis.
3. Determine the statistical significance of *r* when N is 30 or larger.
   a. Compute $z = r\sqrt{N-1}$
   b. Consult a z table.
   c. Reject the null hypothesis if your z value has a probability of occurring that is equal to or less than your level of significance. For a two-tailed test double the probability shown on the table.

SPEARMAN RANK COEFFICIENT (RHO)

Charles Spearman, a British psychologist, is given credit for the first work on the relationship between ranks. His early writings on the subject became known as Spearman's rank-order correlation. (In reality it was Galton, not Spearman, who developed the idea of rank order correlation and it was Pearson who derived the formula.) Anyway, The *Spearman rank coefficient* is referred to as the *Spearman rho* because it is denoted by the Greek letter ρ. It is a nonparametric measure for use with data that are either reduced to ranks or collected in the form of ranks.

In testing the significance of this correlation you are testing the null hypothesis that states there is zero correlation in the population. The requirements for using the Spearman rho are as follows:

1. Ordinal data
2. Two variables
3. Each subject in the study ranked separately on each variable

The formula for the Spearman rho is:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

The terms used to find the Spearman rho have the following meanings: N = number of individuals in the group, D = difference between the ranks in the column labeled R_1 and the column labeled R_2, ρ = Spearman rank coefficient, and ΣD^2 = square each difference and then find the sum.

COEFFICIENT OF CONTINGENCY

This statistic was discussed in Chapter 8.

REFERENCES

Hays, W.L., *Statistics for the Social Sciences*, 2nd ed., Holt, Rinehart & Winston, New York, 1973.

Siegel, S., *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York, 1956.

11 Advanced Topics in Statistics

This chapter is an overview of advanced statistical topics that an experimenter may want to use in the pursuit of finding the appropriate model for understanding and predicting specific outcomes. Because the topics are very complex and the techniques very tedious, the chapter will focus on explaining some of the idiosyncrasies and identifying some of the tests. However, it is assumed that computer software will be used, and therefore no critical tables are provided. (Readers who are interested in table values should consult any statistics book that deals with these topics.)

WHAT ARE DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION?

In attempting to choose an appropriate analytical technique, we sometimes encounter a problem that involves a categorical dependent variable and several metric (measurable) independent variables. For example, we may wish to distinguish good from bad credit risks. If we had a metric measure of credit risk, then we could use multivariate regression. But we may be able to ascertain only if someone is in the good or bad risk category, and this is not the metric type measure required by multivariate regression analysis.

Discriminant analysis and logistic regression are the appropriate statistical techniques when the dependent variable is categorical (nominal or nonmetric) and the independent variables are metric. In many cases, the dependent variable consists of two groups or classifications (for example, male versus female or high versus low). In other instances, more than two groups are involved, such as low, medium, and high classifications. Discriminant analysis is capable of handling either two groups or multiple (three or more) groups. When two classifications are involved, the technique is referred to as two-group discriminant analysis. When three or more classifications are identified, the technique is referred to as *multiple discriminant analysis (MDA).* Logistic regression, also known as logit analysis, is limited in its basic form to two groups, although alternative formulations can handle more than two groups.

Discriminant analysis involves deriving a variate, the linear combination of the two (or more) independent variables that will discriminate best between *a priori*

defined groups. Discrimination is achieved by setting the variate's weights for each variable to maximize the between-group variance relative to the within-group variance. The linear combination for a discriminant analysis, also known as the discriminant function, is derived from an equation that takes the following form:

$$Z_{jk} = \alpha + W_1X_{1k} + W_2X_{2k} + \ldots + W_nX_{nk}$$

where α = intercept, Z_{jk} = discriminant Z score of discriminant function j for object k, W_i = discriminant weight for independent variable i, and X_{ik} = independent variable i for object *k*.

Discriminant analysis is the appropriate statistical technique for testing the hypothesis that the group means of a set of independent variables for two or more groups are equal. To do so, discriminant analysis multiplies each independent variable by its corresponding weight and adds these products together. The result is a single composite discriminant Z score for each individual in the analysis. By averaging the discriminant scores for all the individuals within a particular group, we arrive at the group mean. This group mean is referred to as a *centroid*. When the analysis involves two groups, there are two centroids; with three groups, there are three centroids; and so forth. The centroids indicate the most typical location of any individual from a particular group, and a comparison of the group centroids shows how far apart the groups are along the dimension being tested.

The test for the statistical significance of the discriminant function is a generalized measure of the distance between the group centroids. It is computed by comparing the distributions of the discriminant scores for the groups. If the overlap in the distributions is small, the discriminant function separates the groups well. If the overlap is large, the function is a poor discriminator between the groups. Two distributions of discriminant scores shown in Figure 11.1 further illustrate this concept. The top diagram represents the distributions of discriminant scores for a function that separates the groups well, whereas the lower diagram shows the distributions of discriminant scores on a function that is a relatively poor discriminator between groups A and B. The shaded areas represent probabilities of misclassifying objects from group A into group B.

Multiple discriminant analysis is unique in one characteristic among the dependence relationships of interest here: if there are more than two groups in the dependent variable, discriminant analysis will calculate more than one discriminant function. As a matter of fact, it will calculate NG –1 functions, where NG is the number of groups. Each discriminant function will calculate a discriminant Z score. In the case of a three-group dependent variable, each object will have a score for discriminant functions one and two, allowing the objects to be plotted in two dimensions, with each dimension representing a discriminant function. Thus, discriminant analysis is not limited to a single variate, as is multiple regression, but creates multiple variates representing dimensions of discrimination among the groups.

Logistic regression is a specialized form of regression that is formulated to predict and explain a binary (two-group) categorical variable rather than a metric dependent measure. The form of the logistic regression variate is similar to the variate in multiple

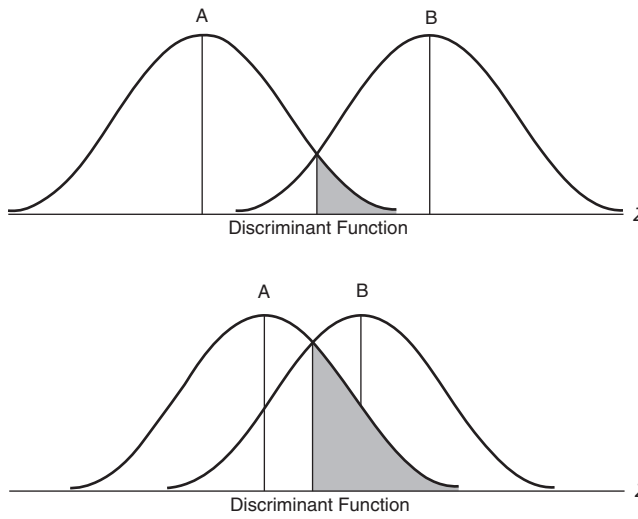


FIGURE 11.1 Univariate representation of discriminant Z scores.

regression. The variate represents a single multivariate relationship with regression-like coefficients that indicate the relative impact of each predictor variable. The differences between logistic regression and discriminant analysis will become more apparent in our discussion of logistic regression's unique characteristics later in this chapter. Yet many similarities also exist between the two methods. When the basic assumptions of both methods are met, they each give comparable predictive and classificatory results and employ similar diagnostic measures. Logistic regression, however, has the advantage of being less affected than discriminant analysis when the basic assumptions, particularly normality of the variables, are not met. It also can accommodate nonmetric variables through dummy-variable coding, just as regression can. Logistic regression is limited, however, to prediction of only a two-group dependent measure. Thus, in cases for which three or more groups form the dependent measure, discriminant analysis is better suited.

Analogy with Regression and MANOVA

The application and interpretation of discriminant analysis is much the same as in regression analysis; that is, the discriminant function is a linear combination (variate) of metric measurements for two or more independent variables and is used to describe or predict a single dependent variable. The key difference is that discriminant analysis is appropriate for research problems in which the dependent variable is categorical (nominal or nonmetric), whereas regression is utilized when the dependent variable is metric. As discussed earlier, logistic regression is a variant of regression, thus having many similarities except for the type of dependent variable. Discriminant analysis is also comparable to "reversing" multivariate analysis of variance (MANOVA). In discriminant analysis, the single dependent variable is categorical, and the independent variables are metric. The opposite is

true of MANOVA, which involves metric dependent variables and categorical independent variable(s).

Discriminant Analysis (DA)

DA was initially developed by Fisher (1936) for the purpose of classifying objects into one of two clearly defined groups. Shortly thereafter, DA was generalized to problems of classification into any number of groups and has been labeled Multiple Discriminant Analysis (MDA). For some time, DA was used exclusively for taxonomic problems in various disciplines (e.g., botany, biology, geology, clinical psychology, vocational guidance). In recent years, DA has come into use as a method of studying group differences on several variables simultaneously. Because of some common features of DA and Multivariate Analysis of Variance (MANOVA), some researchers treat the two as interchangeable methods for studying group differences on multiple variables. More often, however, it is suggested that DA be used after a MANOVA for the purpose of identifying the dimensions along which the groups differ. For a comprehensive review of the various uses of DA see Huberty (1975). Good introductory treatments of DA will be found in Klecka (1980) and Tatsuoka (1970, 1976).

The discussion offered here is limited to the use of DA for the purpose of studying group differences. Sophisticated classification methods, of which DA is but one, are available and are discussed, among others, by Rulon et al. (1967), Overall and Klett (1972), Tatsuoka (1974, 1975), and Van Ryzin (1977).

To understand the DA it is necessary to discuss the concept of Sums of Squares and Cross Products (SSCP) matrices.

SSCP

Whereas in the univariate analysis of variance the total sum of squares of the dependent variable is partitioned into two components: (1) pooled within-groups sum of squares and (2) between-groups sum of squares — with multiple dependent variables it is possible to calculate the within and between groups sums of squares for each of them. In addition, the total sum of cross products between any two variables can be partitioned into (1) pooled within groups sum of products and (2) between-groups sum of products. With multiple dependent variables, it is convenient to assemble the sums of squares and cross products in the following three matrices: **W** = pooled within-groups SSCP; **B** = between-groups SSCP; **T** = total SSCP. To clarify these notions, assume that there are only two dependent variables. Accordingly, the elements of the above matrices are:

$$\mathbf{W} = \begin{bmatrix} SS_{w_1} & SCP_w \\ SCP_w & SS_{w_2} \end{bmatrix}$$

where SS_{w_1} = pooled sum of squares within groups for variable 1, SS_{w_2} = pooled sum of squares within groups for variable 2, and SCP_w = pooled within-groups sum of products of variables 1 and 2.

$$\mathbf{B} = \begin{bmatrix} SS_{b_1} & SCP_b \\ SCP_b & SS_{b_2} \end{bmatrix}$$

where SS_{b_1} and SS_{b_2}, are the between-groups sums of squares for variables 1 and 2, respectively, and SCP_b is the between-groups sum of cross products of variables 1 and 2.

$$\mathbf{T} = \begin{bmatrix} SS_1 & SCP_{12} \\ SCP_{12} & SS_2 \end{bmatrix}$$

where SS_1 and SS_2 are the total sums of squares for variables 1 and 2, respectively, and SCP_{12} is the total sum of cross products of variables 1 and 2. Note that the elements of **T** are calculated as if all the subjects belong to a single group.

Because **T** = **W** + **B**, the elements of the total SSCP matrix (**T**) can be obtained by adding **W** and **B**. This is an important concept and it should be noted that normally **W**, **B**, and **T** are obtained by using matrix operations on the raw score matrices. This is how computer programs are written. Also, as shown above, only two of the three matrices have to be calculated. The third may be obtained by addition or subtraction, whatever the case may be. Thus, **T** was obtained above by adding **W** and **B**. If, instead, **T** and **W** were calculated, then **B** = **T** – **W**, or **W** = **T** – **B**.

Elements of DA

Although the presentation of DA for two groups may be simplified (see, for example, Green, 1978, Chapter 4; Lindeman et al., 1980, Chapter 6), it was felt that it will be more instructive to present the general case — that is, for two groups or more. Therefore, although in the presentation that follows the equations are applied to DA with two groups, the same equations are applicable to DA with any number of groups. Calculation of DA, particularly the eigenvalues, can become very complicated. Consequently, DA is generally calculated by the use of a computer program.

The basic idea of DA is to find a set of weights, v, by which to weight the scores of each individual so that the ratio of **B** (between-groups SSCP) to **W** (pooled within-groups SSCP) is maximized, thereby leading to maximum discrimination among the groups. This may be expressed as follows:

$$\lambda = \frac{\mathbf{v}'\mathbf{B}\mathbf{v}}{\mathbf{v}'\mathbf{W}\mathbf{v}}$$

where $\mathbf{v}'$ and **v** are a row and column vectors of weights, respectively. λ is referred to as the discriminant criterion.

A solution of λ is obtained by solving the following determinantal equation:

$$|\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}| = 0$$

where $\mathbf{W}^{-1}$ is the inverse of **W**, and **I** is an identity matrix. λ is referred to as the largest eigenvalue, or characteristic root, of the matrix, the determinant of which is set equal to zero. With two groups, only one eigenvalue may be obtained. To solve this equation, first, the inverse of **W** has to be calculated (see Appendix A for a review of simple matrix calculations), and the determinant of **W** will be calculated. Second, we proceed to solve for λ so that the determinant of the matrix will be equal to zero.

At this point, having calculated λ, the weights, **v**, are calculated by solving the following:

$$(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I})\mathbf{v} = 0$$

The terms in parenthesis are those used previously in the determinantal equation; **v** is referred to as the eigenvector or the characteristic vector. Using the value of λ and the values of $\mathbf{W}^{-1}\mathbf{B}$ that we obtained earlier, we can proceed to solve the homogeneous equations. The results are coefficients that have a constant proportionality.

MEASURES OF ASSOCIATION

As in the case of univariate analysis, it is desirable to have a measure of association between the independent and the dependent variables in multivariate analysis. Of several such measures that have been proposed (Huberty, 1972, 1975; Shaffer and Gillo, 1974; Smith, 1972; Stevens, 1972; Tatsuoka, 1970, 1971), only one will be presented here. The measure to be presented is related to Wilks' Λ (lambda), which is defined as

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|}$$

where **W** = pooled within-groups SSCP and **T** = total SSCP. Note that lambda is a ratio of the determinants of these two matrices.

Before describing the measure of association that is related to lambda, it will be instructive to show how lambda can be expressed for the case of univariate analysis. Recall that in the univariate analysis of variance the total sum of squares (SS_t) is partitioned into between-groups sum of squares (SS_b) and within-groups sum of squares (SS_w). Accordingly, in univariate analysis,

$$\Lambda = \frac{SS_w}{SS_t}$$

Since $SS_t = SS_b + SS_w$, Λ can also be written

$$\Lambda = \frac{SS_t - SS_b}{SS_t} = 1 - \frac{SS_b}{SS_t}$$

and, from the preceding,

$$\frac{SS_b}{SS_t} = 1 - \Lambda$$

As is well known, the ratio of SS_b to SS_t is defined as η^2 — the proportion of variance of the dependent variable accounted for by the independent variable or group membership. It is clear, then, that lambda indicates the proportion of variance of the dependent variable not accounted for by the independent variable or the proportion of error variance, and that lambda may vary from zero to one. When $\Lambda = 0$ it means that $SS_b = SS_t$, and that the proportion of error variance is equal to zero. When, on the other hand, $\Lambda = 1$, it means that $SS_b = 0$ ($SS_w = SS_t$) and that the proportion of error variance is equal to one.

Finally, when the dependent variable is regressed on coded vectors that represent a categorical independent variable, the following equivalences hold:

$$SS_w = SS_{res}; \quad SS_b = SS_{reg}; \quad \eta^2 = R^2$$

where SS_{res} = residual sum of squares, SS_{reg} = regression sum of squares, and R^2 = squared multiple correlation of the dependent variable with the coded vectors. Accordingly, Λ may be expressed as follows

$$\Lambda = \frac{SS_{res}}{SS_t} = 1 - \frac{SS_{reg}}{SS_t} = 1 - R^2$$

and

$$R^2 = 1 - \Lambda$$

From the above, one may conceive of $1 - \Lambda$ in multivariate analysis as a generalization of η^2 or R^2 of univariate analysis. When in multivariate analysis $\Lambda = 1$ it means that no association exists between the independent and the dependent variables. When on the other hand, $\Lambda = 0$, it means that a perfect association exists between the independent and the dependent variables.

It would be an incomplete discussion if we did not mention the relationship of the test of Λ to the F test. At least in a special case for two groups the following formula is the one that is used:

$$F = [(1 - \Lambda)/t]/[\Lambda/(N - t - 1)]$$

where t is the number of dependent variables and N is the total number of subjects. The df for this F ratio are t and $N - t - 1$. This is identical in form to the test we mentioned above (R^2), when a coded vector representing group membership was regressed on the dependent variables.

A NOTE ON MULTIPLE DISCRIMINANT ANALYSIS

Although the discussion of the equations for DA was applied to the case of two groups, it should be emphasized that it was said earlier that the same equations apply to DA with any number of groups. With more than two groups, more than one discriminant function is calculated. The number of discriminant functions that can be calculated is equal to the number of groups minus one or to the number of dependent variables, whichever is smaller. Thus, with three groups, for example, only two discriminant functions can be calculated, regardless of the number of dependent variables. If, on the other hand, six groups but only three dependent variables are present, the number of discriminant functions that can be calculated is three (the number of the dependent variables).

In the beginning of this chapter, it was mentioned that for the case of two groups, DA can be calculated by multiple regression analysis in which the groups are represented by a coded vector. With more than two groups, it is necessary to use more than one coded vector. Under such circumstances, multiple regression analysis cannot be used; instead, a canonical analysis with coded vectors may be used to calculate DA for any number of groups.

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

MANOVA is an extension of univariate analysis of variance designed to simultaneously test differences among groups on multiple dependent variables. Several tests have been proposed for this purpose. Probably the most widely used among them is a test of Wilks' Λ. That is, a test of Λ (shown below) serves as an overall test of the null hypothesis of the equality of mean vectors of two or more groups.

In an earlier section, Λ was discussed in detail in the context of DA, though the test of significance for Λ was not shown. A question that naturally arises is: Since Λ may be obtained in both DA and MANOVA, in what way do these approaches differ? Issues of classification to which DA but not MANOVA may be applied aside, it was said earlier that some researchers treat MANOVA and DA as interchangeable when the concern is the study of group differences. But other researchers recommend that MANOVA be applied first in order to determine whether there are overall significant differences among the groups. This is accomplished by testing Λ. If the null hypothesis is rejected, it is recommended that DA be used to identify the variables on which the groups differ to the greatest extent and the nature of the dimensions on which they differ.

It was stated above that when more than two groups are being studied, more than one discriminant function is obtained. In such situations, the test of Λ still refers to overall differences among the groups. But if it is found that Λ is statistically significant, it may turn out that only one or two discriminant functions are statistically significant, even though the data allow the calculation of a greater number of such functions.

Testing the Assumptions of Multivariate Analysis

As we already have discussed several times, data are the life blood of any statistical analysis. However, the data must be appropriate and applicable for the type of

analysis that we want to conduct. The introduction to this book mentioned some of the preliminary steps in preparing data for analysis. Here we are going to address the final step in examining the data, which involves testing the assumptions underlying multivariate analysis. The need to test the statistical assumptions is increased in multivariate applications because of two characteristics of multivariate analysis. First, the complexity of the relationships, owing to the typical use of a large number of variables, makes the potential distortions and biases more potent when the assumptions are violated. This is particularly true when the violations compound to become even more detrimental than if considered separately. Second, the complexity of the analyses and of the results may mask the "signs" of assumption violations apparent in the simpler univariate analyses. In almost all instances, the multivariate procedures will estimate the multivariate model and produce results even when the assumptions are severely violated. Thus, the experimenter must be aware of any assumption violations and the implications they may have for the estimation process or the interpretation of the results.

Assessing Individual Variables Versus the Variate

Multivariate analysis requires that the assumptions underlying the statistical techniques be tested twice: first for the separate variables, akin to the tests of assumption for univariate analyses, and second for the multivariate model variate, which acts collectively for the variables in the analysis and thus must meet the same assumptions as individual variables do.

Normality

The most fundamental assumption in multivariate analysis is normality, referring to the shape of the data distribution for an individual metric variable and its correspondence to the normal distribution, the benchmark for statistical methods. If the variation from the normal distribution is sufficiently large, all resulting statistical tests are invalid because normality is required for use of the F and *t* statistics. Both the univariate and the multivariate statistical methods discussed in this volume are based on the assumption of univariate normality, with the multivariate methods also assuming multivariate normality. Univariate normality for a single variable is easily tested, and a number of corrective measures are possible, as shown later. In a simple sense, multivariate normality (the combination of two or more variables) means that the individual variables are normal in a univariate sense and that their combinations are also normal. Thus, if a variable is multivariate normal, it is also univariate normal. However, the reverse is not necessarily true (two or more univariate normal variables are not necessarily multivariate normal). Thus a situation in which all variables exhibit univariate normality will help gain, although not guarantee, multivariate normality. Multivariate normality is more difficult to test, but some tests are available for situations in which the multivariate technique is particularly affected by a violation of this assumption. In this text, we focus on assessing and achieving univariate normality for all variables and address multivariate normality only when it is especially critical. Even though large sample sizes tend to diminish the detrimental effects

of nonnormality, the researcher should assess the normality for all variables included in the analysis.

Graphical Analyses of Normality

The simplest diagnostic test for normality is a visual check of the histogram that compares the observed data values with a distribution approximating the normal distribution. Although appealing because of its simplicity, this method is problematic for smaller samples, where the construction of the histogram (e.g., the number of categories or the width of categories) can distort the visual portrayal to such an extent that the analysis is useless. A more reliable approach is the normal probability plot, which compares the cumulative distribution of actual data values with the cumulative distribution of a normal distribution. The normal distribution forms a straight diagonal line, and the plotted data values are compared with the diagonal. If a distribution is normal, the line representing the actual data distribution closely follows the diagonal.

Figure 11.2 shows several normal probability plots and the corresponding univariate distribution of the variable. One characteristic of the distribution's shape, the kurtosis, is reflected in the normal probability plots. Kurtosis refers to the

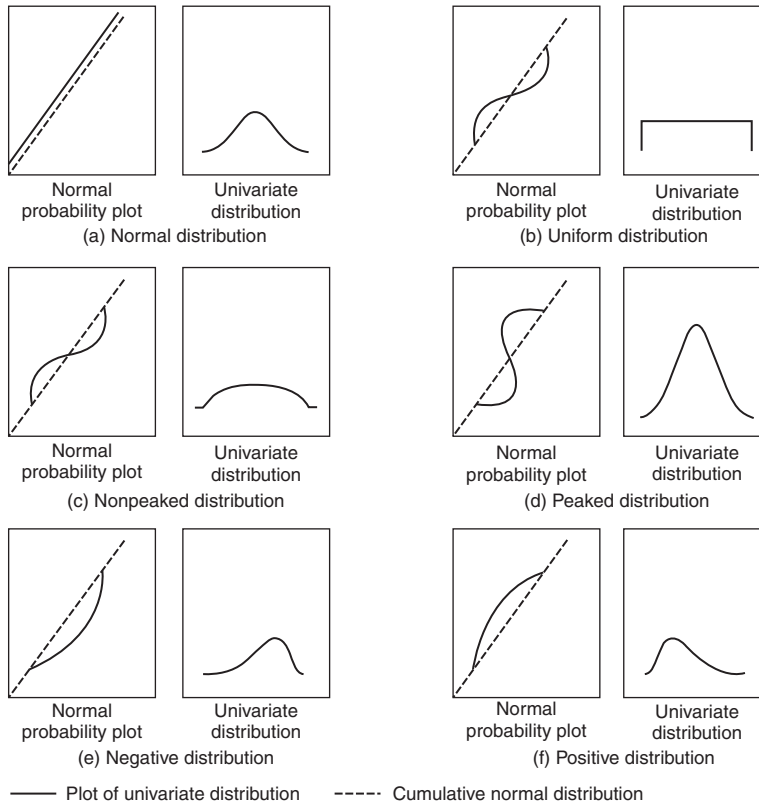


FIGURE 11.2 Normal probability plots and corresponding univariate distributions.

"peakedness" or "flatness" of the distribution compared with the normal distribution. When the line falls below the diagonal, the distribution is flatter than expected. When it goes above the diagonal, the distribution is more peaked than the normal curve. For example, in the normal probability plot of a peaked distribution (Figure 11.2d), we see a distinct S-shaped curve. Initially the distribution is flatter, and the plotted line falls below the diagonal. Then the peaked part of the distribution rapidly moves the plotted line above the diagonal, and eventually the line shifts to below the diagonal again as the distribution flattens. A nonpeaked distribution has the opposite pattern (Figure 11.2c). Another common pattern is a simple arc, either above or below the diagonal, indicating the skewness of the distribution. A negative skewness (Figure 11.2e) is indicated by an arc below the diagonal, whereas an arc above the diagonal represents a positively skewed distribution (Figure 11.2f). An excellent source for interpreting normal probability plots, showing the various patterns and interpretations, is Daniel and Wood (1980). These specific patterns not only identify nonnormality but also tell us the form of the original distribution and the appropriate remedy to apply.

Statistical Tests of Normality

In addition to examining the normal probability plot, one can also use statistical tests to assess normality. A simple test is a rule of thumb based on the skewness and kurtosis values (available as part of the basic descriptive statistics for a variable computed by all statistical programs). The statistic value (z) for the skewness value is calculated as:

$$Z_{\text{skewness}} = \frac{\text{Skewness}}{\sqrt{\frac{6}{N}}}$$

where N is the sample size. A z value can also be calculated for the kurtosis value using the following formula:

$$Z_{\text{kurtosis}} = \frac{\text{Kurtosis}}{\sqrt{\frac{24}{N}}}$$

If the calculated z value exceeds a critical value, then the distribution is nonnormal in terms of that characteristic. The critical value is from a z distribution, based on the significance level we desire. For example, a calculated value exceeding ±2.58 indicates we can reject the assumption about the normality of the distribution at the .01 probability level. Another commonly used critical value is *t* 1.96, which corresponds to a .05 error level. (Specific statistical tests are also available in SPSS, SAS, BMDP, and most other programs. The two most common are the Shapiro-Wilks test and a modification of the Kolmogorov-Smirnov test. Each calculates the level of significance for the differences from a normal distribution. The experimenter should always remember that tests of significance are less useful in small samples [fewer

than 30] and quite sensitive in large samples [exceeding 1,000 observations]. Thus, the researcher should always use both the graphical plots and any statistical tests to assess the actual degree of departure from normality.)

Remedies for Nonnormality

A number of data transformations available to accommodate nonnormal distributions are discussed later in the chapter. This chapter confines the discussion to univariate normality tests and transformations. However, when we examine other multivariate methods, such as multivariate regression or multivariate analysis of variance, we discuss tests for multivariate normality as well. Moreover, many times when nonnormality is indicated, it is actually the result of other assumption violations; therefore, remedying the other violations eliminates the nonnormality problem. For this reason, the researcher should perform normality tests after or concurrently with analyses and remedies for other violations. (Those interested in multivariate normality should see Johnson and Wichern [1982] and Weisberg [1985].)

Homoscedasticity

Homoscedasticity is an assumption related primarily to dependence relationships between variables. It refers to the assumption that dependent variable(s) exhibit equal levels of variance across the range of predictor variable(s). Homoscedasticity is desirable because the variance of the dependent variable being explained in the dependence relationship should not be concentrated in only a limited range of the independent values. Although the dependent variables must be metric, this concept of an equal spread of variance across independent variables can be applied when the independent variables are either metric or nonmetric. With metric independent variables, the concept of homoscedasticity is based on the spread of dependent variable variance across the range of independent variable values, which is encountered in techniques such as multiple regression. The same concept also applies when the independent variables are nonmetric. In these instances, such as are found in ANOVA and MANOVA, the focus now becomes the equality of the variance (single dependent variable) or the variance/covariance matrices (multiple independent variables) across the groups formed by the nonmetric independent variables. The equality of variance/covariance matrices is also seen in discriminant analysis, but in this technique the emphasis is on the spread of the independent variables across the groups formed by the nonmetric dependent measure. In each of these instances, the purpose is the same: to ensure that the variance used in explanation and prediction is distributed across the range of values, thus allowing for a "fair test" of the relationship across all values of the nonmetric variables.

In most situations, we have many different values of the dependent variable at each value of the independent variable. For this relationship to be fully captured, the dispersion (variance) of the dependent variable values must be equal at each value of the predictor variable. Most problems with unequal variances stem from one of two sources. The first source is the type of variables included in the model. For example, as a variable increases in value (e.g., units ranging from near zero to millions), there is naturally a wider range of possible answers for the larger values.

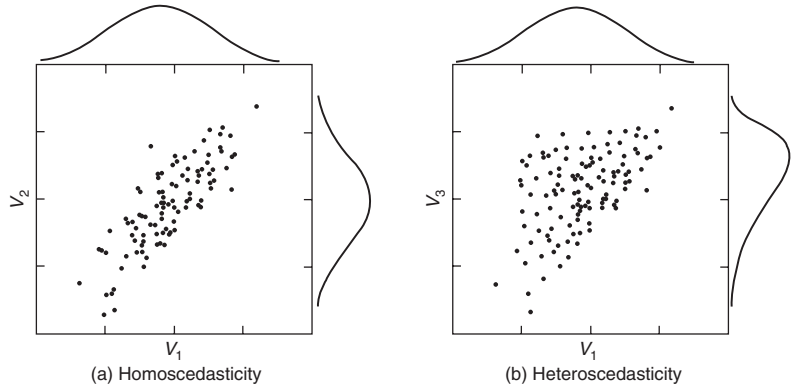


FIGURE 11.3 Scatterplots of homoscedastic and heteroscedastic relationships.

The second source results from a skewed distribution that creates heteroscedasticity, In Figure 11.3a, the scatterplots of data points for two variables (V_1 and V_2) with normal distributions exhibit equal dispersion across all data values (i.e., homoscedasticity). However, in Figure 11.3b, we see unequal dispersion (heteroscedasticity) caused by skewness of one of the variables (V_3). For the different values of V_3, there are different patterns of dispersion for V_1. This will cause the predictions to be better at some levels of the independent variable than at others. Violating this assumption often makes hypothesis tests either too conservative or too sensitive.

The effect of heteroscedasticity is also often related to sample size, especially when examining the variance dispersion across groups. For example, in ANOVA or MANOVA, the impact of heteroscedasticity on the statistical test depends on the sample sizes associated with the groups of smaller and larger variances. In multiple regression analysis, similar effects would occur in highly skewed distributions where there were disproportionate numbers of respondents in certain ranges of the independent variable.

Graphical Tests of Equal Variance Dispersion

The test of homoscedasticity for two metric variables is best examined graphically. The most common application of this form of assessment occurs in multiple regression, which is concerned with the dispersion of the dependent variable across the values of the metric independent variables. Because the focus of regression analysis is on the regression variate, the graphical plot of residuals is used to reveal the presence of homoscedasticity (or its opposite, heteroscedasticity). The earlier discussion of residual analysis details these procedures. Boxplots work well to represent the degree of variation between groups formed by a categorical variable. The length of the box and the whiskers each portray the variation of data within that group. A typical boxplot is shown in Figure 11.4.

Statistical Tests for Homoscedasticity

The statistical tests for equal variance dispersion relate to the variances within groups formed by nonmetric variables. The most common test, the Levene test, can be used to assess whether the variances of a single metric variable are equal across any

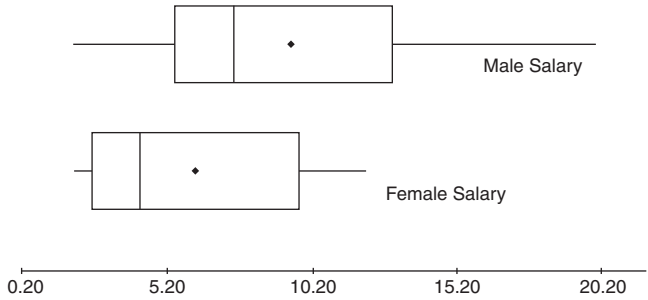


FIGURE 11.4 A typical comparison of side-by-side boxplots.

number of groups. If more than one metric variable is being tested, so that the comparison involves the equality of variance/covariance matrices, the Box's M test is applicable. The Box's M is a statistical test for the equality of the covariance matrices of the independent variables across the groups of the dependent variable. If the statistical significance is greater than the critical level (.01), then the equality of the covariance matrices is supported. If the test shows statistical significance, then the groups are deemed different and the assumption is violated. The Box's M test is available in both multivariate analysis of variance and discriminant analysis.

Remedies for Heteroscedasticity

Heteroscedastic variables can be remedied through data transformations similar to those used to achieve normality. As mentioned earlier, many times heteroscedasticity is the result of nonnormality of one of the variables, and correction of the nonnormality also remedies the unequal dispersion of variance.

Linearity

An implicit assumption of all multivariate techniques based on correlational measures of association, including multiple regression, logistic regression, factor analysis, and structural equation modeling, is linearity. Because correlations represent only the linear association between variables, nonlinear effects will not be represented in the correlation value. This results in an underestimation of the actual strength of the relationship. It is always prudent to examine all relationships to identify any departures from linearity that may impact the correlation.

Identifying Nonlinear Relationships

The most common way to assess linearity is to examine scatterplots of the variables and to identify any nonlinear patterns in the data. An alternative approach is to run a simple regression analysis and to examine the residuals. The residuals reflect the unexplained portion of the dependent variable; thus, any nonlinear portion of the relationship will show up in the residuals. The examination of residuals can also be applied to multiple regression, where the researcher can detect any nonlinear effects not represented in the regression variate. If a nonlinear relationship is detected, the most direct approach is to transform one or both variables to achieve linearity.

WHAT IS FACTOR ANALYSIS?

Factor analysis is a generic name given to a class of multivariate statistical methods whose primary purpose is to define the underlying structure in a data matrix. Broadly speaking, it addresses the problem of analyzing the structure of the interrelationships (correlations) among a large number of variables (e.g., test scores, test items, questionnaire responses) by defining a set of common underlying dimensions, known as factors. With factor analysis, the experimenter can first identify the separate dimensions of the structure and then determine the extent to which each variable is explained by each dimension. Once these dimensions and the explanation of each variable are determined, the two primary uses for factor analysis — summarization and data reduction — can be achieved. In summarizing the data, factor analysis derives underlying dimensions that, when interpreted and understood, describe the data in a much smaller number of concepts than the original individual variables. Data reduction can be achieved by calculating scores for each underlying dimension and substituting them for the original variables.

We introduce factor analysis as our first multivariate technique because it can play a unique role in the application of other multivariate techniques. As already discussed, the primary advantage of multivariate techniques is their ability to accommodate multiple variables in an attempt to understand the complex relationships not possible with univariate and bivariate methods. Increasing the number of variables also increases the possibility that the variables are not all uncorrelated and representative of distinct concepts. Instead, groups of variables may be interrelated to the extent that they are all representative of a more general concept. This may be by design, such as the attempt to measure the many facets of personality or store image, or may arise just from the addition of new variables. In either case, the researcher must know how the variables are interrelated to better interpret the results. Finally, if the number of variables is too large or there is a need to better represent a smaller number of concepts rather than many facets, factor analysis can assist in selecting a representative subset of variables or even creating new variables as replacements for the original variables while still retaining their original character.

Factor analysis differs from the dependence techniques discussed in the next section (i.e., multiple regression, discriminant analysis, multivariate analysis of variance, or canonical correlation), in which one or more variables are explicitly considered the criterion or dependent variables and all others are the predictor or independent variables. Factor analysis is an interdependence technique in which all variables are simultaneously considered, each related to all others, and still employing the concept of the variate, the linear composite of variables. In factor analysis, the variates (factors) are formed to maximize their explanation of the entire variable set, not to predict a dependent variable(s). If we were to draw an analogy to dependence techniques, it would be that each of the observed (original) variables is a dependent variable that is a function of some underlying and latent set of factors (dimensions) that are themselves made up of all other variables. Thus, each variable is predicted by all others. Conversely, one can look at each factor (variate) as a dependent variable that is a function of the entire set of observed variables. Either

analogy illustrates the differences in purpose between dependence (prediction) and interdependence (identification of structure) techniques.

Factor analytic techniques can achieve their purposes from either an exploratory or confirmatory perspective. There is continued debate concerning the appropriate role for factor analysis. Many researchers consider it only exploratory — useful in searching for structure among a set of variables or as a data reduction method. From this perspective, factor analytic techniques "take what the data give you" and do not set any *a priori* constraints on the estimation of components or the number of components to be extracted. For many if not most applications, this use of factor analysis is appropriate. However, in other situations, the experimenter has preconceived thoughts on the actual structure of the data, based on theoretical support or prior research. The experimenter may wish to test hypotheses involving issues such as which variables should be grouped together on a factor or the precise number of factors. In these instances, the experimenter requires that factor analysis take a confirmatory approach — that is, assess the degree to which the data meet the expected structure.

MULTIPLE REGRESSION ANALYSIS

Multiple regression analysis is a statistical technique that can be used to analyze the relationship between a single dependent (criterion) variable and several independent (predictor) variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the single dependent value selected by the experimenter. Each independent variable is weighted by the regression analysis procedure to ensure maximal prediction from the set of independent variables. The weights denote the relative contribution of the independent variables to the overall prediction and facilitate interpretation as to the influence of each variable in making the prediction, although correlation among the independent variables complicates the interpretive process. The set of weighted independent variables forms the regression variate, a linear combination of the independent variables that best predicts the dependent variable. The regression variate, also referred to as the regression equation or regression model, is the most widely known example of a variate among the multivariate techniques.

Multiple regression analysis is a dependence technique. Because of this, you as an experimenter must be able to classify the variables as dependent and independent. However, because regression is also a statistical tool, it should be used when the variables are metric. Only under certain circumstances it is possible to include nonmetric data. When we do that, appropriate transformation of data must occur.

Representing Curvilinear Effects with Polynomials

Several types of data transformations are appropriate for linearizing a curvilinear relationship. Direct approaches involve modifying the values through some arithmetic transformation (e.g., taking the square root or logarithm of the variable). However, such transformations have several limitations. First, they are helpful only in a simple curvilinear relationship (a relationship with only one turning or inflection

point). Second, they do not provide any statistical means for assessing whether the curvilinear or linear model is more appropriate. Finally, they accommodate only univariate relationships and not the interaction between variables when more than one independent variable is involved. We now discuss a means of creating new variables to explicitly model the curvilinear components of the relationship and address each of the limitations inherent in data transformations.

Polynomials are power transformations of an independent variable that add a nonlinear component for each additional power of the independent variable. The power of 1 (X^1) represents the linear component and is the simplest form representing a line. The power of 2, the variable squared (X^2), represents the quadratic component. In graphical terms, X^2 represents the first inflection point. A cubic component, represented by the variable cubed (X^3), adds a second inflection point. With these variables and even higher powers, we can accommodate more complex relationships than are possible with only transformations. For example, in a simple regression model, a curvilinear model with one turning point can be modeled with the equation

$$Y = b_0 + b_1X_1 + b_2X_1^2$$

where b_0 = intercept, b_1X_1 = linear effect of X_1, and $b_2X_1^2$ = curvilinear effect of X_1.

Although any number of nonlinear components may be added, the cubic term is usually the highest power used. As each new variable is entered into the regression equation, we can also perform a direct statistical test of the nonlinear components, which we cannot do with data transformations. Three (two nonlinear and one linear) relationships are shown in Figure 11.5. For interpretation purposes, the positive quadratic term indicates a U-shaped curve, whereas a negative coefficient indicates a $\cap$-shaped curve.

Multivariate polynomials are created when the regression equation contains two or more independent variables. We follow the same procedure for creating the polynomial terms as before but must also create an additional term, the interaction term (X_1X_2), which is needed for each variable combination to represent fully the multivariate effects. In graphical terms, a two-variable multivariate polynomial is portrayed

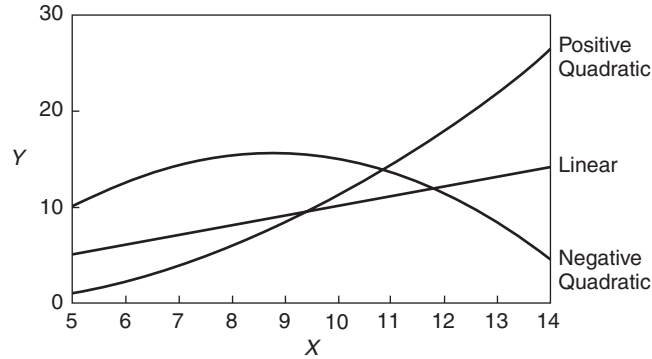


FIGURE 11.5 Representing nonlinear relationships with polynomials.

by a surface with one peak or valley. For higher-order polynomials, the best form of interpretation is obtained by plotting the surface from the predicted values.

How many terms should be added? Common practice is to start with the linear component and then sequentially add higher-order polynomials until nonsignificance is achieved. The use of polynomials, however, also has potential problems. First, each additional term requires a degree of freedom, and this may be particularly restrictive with small sample sizes. This limitation does not occur with data transformation. Also, multicollinearity is introduced by the additional terms and makes statistical significance testing of the polynomial terms inappropriate. Instead, the experimenter must compare the R^2 values from the equation model with linear terms to the R^2 for the equation with the polynomial terms. Testing for the statistical significance of the incremental R^2 is the appropriate manner of assessing the impact of the polynomials.

Standardizing the Regression Coefficients: Beta Coefficients

If each of our independent variables had been standardized before we estimated the regression equation, we would have found different regression coefficients. The coefficients resulting from standardized data are called beta (β) coefficients. Their advantage is that they eliminate the problem of dealing with different units of measurement, thus reflecting the relative impact on the dependent variable of a change in one standard deviation in either variable. Now that we have a common unit of measurement, we can determine which variable has the most impact.

Three cautions must be observed when using beta coefficients. First, they should be used as a guide to the relative importance of individual independent variables only when collinearity is minimal. Second, the beta values can be interpreted only in the context of the other variables in the equation. For example, a beta value for family size reflects its importance only in relation to family income, not in any absolute sense. If another independent variable were added to the equation, the beta coefficient for family size would probably change, because there would likely be some relationship between family size and the new independent variable. The third caution is that the levels (e.g., families of size five, six, and seven persons) affect the beta value. Had we found families of size eight, nine, and ten, the value of beta would likely change. In summary, beta coefficients should be used only as a guide to the relative importance of the independent variables included in the equation, and only over the range of values for which sample data actually exist.

Assessing Multicollinearity

A key issue in interpreting the regression variate is the correlation among the independent variables. This is a data problem, not a problem of model specification. The ideal situation for an experimenter would be to have a number of independent variables highly correlated with the dependent variable, but with little correlation among themselves. Yet in most situations, particularly situations involving consumer response data, there will be some degree of multicollinearity. In some other occasions, such as using dummy variables to represent nonmetric variables or polynomial terms for nonlinear effects, the researcher is creating situations of high

multicollinearity. The researcher's task is to assess the degree of multicollinearity and determine its impact on the results and the necessary remedies if needed. In the following sections we discuss the effects of multicollinearity and then detail some useful diagnostic procedures and possible remedies.

The effects of multicollinearity can be categorized in terms of *explanation* and *estimation.* The effects on explanation primarily concern the ability of the regression procedure and the experimenter to represent and understand the effects of each independent variable in the regression variate. As multicollinearity occurs (even at the relatively low levels of .30 or so), the process for separating the effects of individuals becomes more difficult. First, it limits the size of the coefficient of determination and makes it increasingly more difficult to add unique explanatory prediction from additional variables. Second, and just as important, it makes determining the contribution of each independent variable difficult because the effects of the independent variables are "mixed" or confounded. Multicollinearity results in larger portions of shared variance and lower levels of unique variance from which the effects of the individual independent variables can be determined. For example, assume that one independent variable (X_1) has a correlation of .60 with the dependent variable, and a second independent variable (X_2) has a correlation of .50. Then X_1 would explain 36% (obtained by squaring the correlation of .60) of the variance of the dependent variable, and X_2 would explain 25% (correlation of .50 squared). If the two independent variables are not correlated with each other at all, there is no "overlap," or sharing, of their predictive power. The total explanation would be their sum, or 61%. But as collinearity increases, there is some "sharing" of predictive power, and the collective predictive power of the independent variables decreases.

Figure 11.6 portrays the proportions of shared and unique variance for our example of two independent variables in varying instances of collinearity. If the collinearity of these variables is zero, then the individual variables predict 36 and 25% of the variance in the dependent variable, for an overall prediction (R^2) of 61%. But as multicollinearity increases, the total variance explained decreases. Moreover, the amount of unique variance for the independent variables is reduced to levels that make estimation of their individual effects quite problematic.

In addition to the effects on explanation, multicollinearity can have substantive effects on the estimation of the regression coefficients and their statistical significance tests. First, the extreme case of multicollinearity in which two or more variables are perfectly correlated, termed singularity, prevents the estimation of any coefficients. In this instance, the singularity must be removed before the estimation of coefficients can proceed. Even if the multicollinearity is not perfect, high degrees of multicollinearity can result in regression coefficients being incorrectly estimated and even having the wrong signs.

Because of these potential problems, the effects of multicollinearity can be substantial. In any regression analysis, the assessment of multicollinearity should be undertaken in two steps: identification of the extent of collinearity, and assessment of the degree to which the estimated coefficients are affected. If corrective action is needed, assess the correlation matrix for the independent variables first and then follow with both pairwise and multiple-variable collinearity. Two of the most

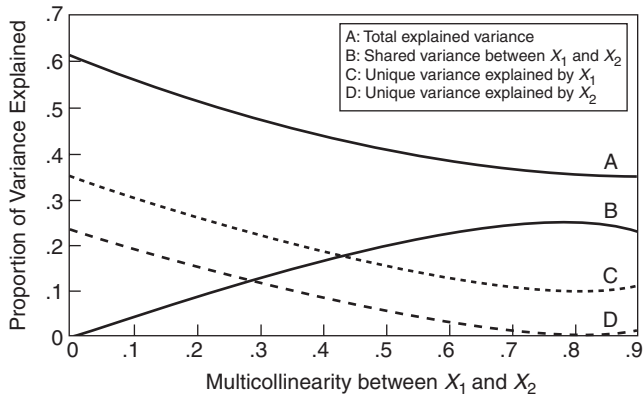


FIGURE 11.6 Proportions of unique and shared variance by levels of multicollinearity.

common measures are the tolerance value and its inverse — the variance inflation factor. These measures tell us the degree to which each independent variable is explained by the other independent variable. (Tolerance is the amount of variability of the selected independent variable not explained by the other independent variables. A common cut-off threshold is a tolerance of .10. However, each study should be evaluated on its own merits and appropriately evaluated.)

WHAT IS MULTIVARIATE ANALYSIS OF VARIANCE?

Multivariate analysis of variance is the multivariate extension of the univariate techniques for assessing the differences between group means. The univariate procedures include the *t* test for two-group situations and ANOVA for situations with three or more groups defined by two or more independent variables. Before proceeding with our discussion of the unique aspects of MANOVA, let us review the basic principles of the univariate techniques.

Univariate Procedures for Assessing Group Differences

These procedures are classified as univariate not because of the number of independent variables, but instead because of the number of dependent variables. In multiple regression, the terms univariate and multivariate refer to the number of independent variables, but for ANOVA and MANOVA, the terminology applies to the use of single or multiple dependent variables. The following discussion addresses the two most common types of univariate procedures, the *t* test, which compares a dependent variable across two groups, and ANOVA, which is used whenever the number of groups is three or more.

The *t* Test

The *t* test assesses the statistical significance of the difference between two independent sample means. For example, an experimenter may expose two groups of

respondents to different advertisements reflecting different advertising messages — one informational and one emotional — and subsequently ask each group about the appeal of the message on a 10-point scale, with 1 being poor and 10 being excellent. The two different advertising messages represent a treatment with two levels (informational versus emotional). A treatment, also known as a factor, is a nonmetric independent variable, experimentally manipulated or observed, that can be represented in various categories or levels. In our example, the treatment is the effect of emotional versus informational appeals.

To determine whether the two messages are viewed differently (meaning that the treatment has an effect), a *t* statistic is calculated. The *t* statistic is the ratio of the difference between the sample means ($\mu_1 - \mu_2$) to their standard error. The standard error is an estimate of the difference between means to be expected because of sampling error, rather than real differences between means. This can be shown in the equation

$$t \text{ statistic} = \frac{\mu_1 - \mu_2}{SE\mu_1\mu_2}$$

where μ_1 = mean of group 1, μ_2 = mean of group 2, and $SE\mu_1\mu_2$ = standard error of the difference in group means.

By forming the ratio of the actual difference between the means to the difference expected due to sampling error, we quantify the amount of the actual impact of the treatment that is due to random sampling error. In other words, the *t* value, or *t* statistic, represents the group difference in terms of standard errors. If the *t* value is sufficiently large, then statistically we can say that the difference was not due to sampling variability but represents a true difference. This is done by comparing the *t* statistic to the critical value of the *t* statistic (t_{crit}) If the absolute value of the *t* statistic is greater than the critical value, this leads to rejection of the null hypothesis of no difference in the appeals of the advertising messages between groups. This means that the actual difference due to the appeals is statistically larger than the difference expected from sampling error. We determine the critical value (t_{crit}) for our *t* statistic and test the statistical significance of the observed differences by the following procedure:

1. Compute the *t* statistic as the ratio of the difference between sample means and their standard error.
2. Specify a Type I error level (denoted as α, or significance level), which indicates the probability level the experimenter will accept in concluding that the group means are different when in fact they are not.
3. Determine the critical value (t_{crit}) by referring to the *t* distribution with $N_1 + N_2 - 2$ degrees of freedom and a specified α, where N_1 and N_2 are sample sizes.
4. If the absolute value of the computed *t* statistic exceeds t_{crit}, the experimenter can conclude that the two advertising messages have different levels of appeal (i.e., $\mu_1 \neq \mu_2$), with a Type I error probability of α. The

researcher can then examine the actual mean values to determine which group is higher on the dependent value.

Analysis of Variance

In our example for the *t* test, an experimenter exposed two groups of respondents to different advertising messages and subsequently asked them to rate the appeal of the advertisements on a 10-point scale. Suppose we were interested in evaluating three advertising messages rather than two. Respondents would be randomly assigned to one of three groups, and we would have three sample means to compare. To analyze these data, we might be tempted to conduct separate *t* tests for the difference between each pair of means (i.e., group 1 versus group 2; group 1 versus group 3; and group 2 versus group 3).

However, multiple *t* tests inflate the overall Type I error rate. ANOVA avoids this Type I error inflation due to making multiple comparisons of treatment groups by determining in a single test whether the entire set of sample means suggests that the samples were drawn from the same general population. That is, ANOVA is used to determine the probability that differences in means across several groups are due solely to sampling error.

The logic of an ANOVA test is fairly straightforward. As the name "analysis of variance" implies, two independent estimates of the variance for the dependent variable are compared, one that reflects the general variability of respondents within the groups (MS_w) and another that represents the differences between groups attributable to the treatment effects (MS_B):

1. *Within-groups estimate of variance* (MS_w: mean square within groups): This is an estimate of the average random respondent variability on the dependent variable within a treatment group and is based on deviations of individual scores from their respective group means. MS_w is comparable to the standard error between two means calculated in the t test as it represents variability within groups. The value MS_w is sometimes referred to as the error variance.
2. *Between-groups estimate of variance* (MS_B: mean square between groups): The second estimate of variance is the variability of the treatment group means on the dependent variable. It is based on deviations of group means from the overall grand mean of all scores. Under the null hypothesis of no treatment effects (i.e., $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$), this variance estimate, unlike MS_w, reflects any treatment effects that exist; that is, differences in treatment means increase the expected value of MS_B.

Given that the null hypothesis of no group differences is true, MS_W and MS_B represent independent estimates of population variance. Therefore, the ratio of MS_B to MS_w is a measure of how much variance is attributable to the different treatments versus the variance expected from random sampling. The ratio of MS_B to MS_W gives us a value for the F statistic. This is similar to the calculation of the *t* value and can be shown as

$$\text{F statistic} = \frac{MS_b}{MS_w}$$

Because group differences tend to inflate MS_B, large values of the F statistic lead to rejection of the null hypothesis of no difference in means across groups. If the analysis has several different treatments (independent variables), then estimates of MS_B are calculated for each treatment and F statistics are calculated for each treatment. This allows for the separate assessment of each treatment.

To determine if the F statistic is sufficiently large to support rejection of the null hypothesis, follow a process similar to the *t* test. First, determine the critical value for the F statistic (F_{crit}) by referring to the F distribution with $(k - 1)$ and $(N - k)$ degrees of freedom for a specified level of a (where $N = N_1, +... + N_k$ and k = number of groups). If the value of the calculated F statistic exceeds F_{crit}, conclude that the means across all groups are not all equal.

Examination of the group means then allows the experimenter to assess the relative standing of each group on the dependent measure. Although the F statistic test assesses the null hypothesis of equal means, it does not address the question of which means are different. For example, in a three-group situation, all three groups may differ significantly, or two may be equal but differ from the third. To assess these differences, the experimenter can employ either planned comparisons or post hoc tests. We examine some of these methods in a later section.

Multivariate Analysis of Variance

As statistical inference procedures, both the univariate techniques (*t* test and ANOVA) and MANOVA are used to assess the statistical significance of differences between groups. In the *t* test and ANOVA, the null hypothesis tested is the equality of dependent variable means across groups. In MANOVA, the null hypothesis tested is the equality of vectors of means on multiple dependent variables across groups. In the univariate case, a single dependent measure is tested for equality across the groups. In the multivariate case, a variate is tested for equality. In MANOVA, the experimenter actually has two variates, one for the dependent variables and another for the independent variables. The dependent variable variate is of more interest because the metric dependent measures can be combined in a linear combination, as we have already seen in multiple regression and discriminant analysis. The unique aspect of MANOVA is that the variate optimally combines the multiple dependent measures into a single value that maximizes the differences across groups.

The Two-Group Case: Hotelling's T^2

In our earlier univariate example, experimenters were interested in the appeal of two advertising messages. But what if they also wanted to know about the purchase intent generated by the two messages? If only univariate analyses were used, the experimenters would perform separate *t* tests on the ratings of both the appeal of the messages and the purchase intent generated by the messages. Yet the two measures are interrelated; thus, what is really desired is a test of the differences between

the messages on both variables collectively. This is where Hotelling's T^2, a specialized form of MANOVA that is a direct extension of the univariate *t* test, can be used.

Hotelling's T^2 provides a statistical test of the variate formed from the dependent variables that produces the greatest group difference. It also addresses the problem of "inflating" the Type I error rate that arises when making a series of *t* tests of group means on several dependent measures. It controls this inflation of the Type I error rate by providing a single overall test of group differences across all dependent variables at a specified α level.

The computational formula for Hotelling's T^2 represents the results of mathematical derivations used to solve for a maximum *t* statistic (and, implicitly, the most discriminating linear combination of the dependent variables). This is equivalent to saying that if we can find a discriminant function for the two groups that produces a significant T^2, the two groups are considered different across the mean vectors.

How does Hotelling's T^2 provide a test of the hypothesis of no group difference on the vectors of mean scores? Just as the *t* statistic follows a known distribution under the null hypothesis of no treatment effect on a single dependent variable, Hotelling's T^2 follows a known distribution under the null hypothesis of no treatment effect on any of a set of dependent measures. This distribution turns out to be an F distribution with p and $N_1 + N_2 - 2 - 1$ degrees of freedom after adjustment (where p = the number of dependent variables). To get the critical value for Hotelling's T^2, we find the tabled value for F_{crit} at a specified α level and compute T^2_{crit} as follows:

$$T^2_{crit} = [p(N_1 + N_2 - 2)/N_1 + N_2 - p - 1] \times F_{crit}$$

Differences between MANOVA and Discriminant Analysis

So far we have discussed the basic elements of both the univariate and multivariate tests for assessing differences between groups on one or more dependent variables. In doing so, we noted the calculation of the discriminant function, which in the case of MANOVA is the variate of dependent variables that maximizes the difference between groups. The question may arise: What is the difference between MANOVA and discriminant analysis? In some aspects, MANOVA and discriminant analysis are "mirror images." The dependent variables in MANOVA (a set of metric variables) are the independent variables in discriminant analysis, and the single nonmetric dependent variable of discriminant analysis becomes the independent variable in MANOVA. Moreover, both use the same methods in forming the variates and assessing the statistical significance between groups.

The differences, however, center on the objectives of the analyses and the role of the nonmetric variable(s). Discriminant analysis employs a single nonmetric variable as the dependent variable. The categories of the dependent variable are assumed as given, and the independent variables are used to form variates that maximally differ between the groups formed by the dependent variable categories. In MANOVA, the metric variables now act as the dependent variables and the objective becomes finding groups of respondents that exhibit differences on the set of dependent variables. The groups of respondents are not prespecified; instead, the

experimenter uses one or more independent variables (nonmetric variables) to form groups. MANOVA, even while forming these groups, still retains the ability to assess the impact of each nonmetric variable separately.

WHAT IS CONJOINT ANALYSIS?

Conjoint analysis is a multivariate technique used specifically to understand how respondents develop preferences for products or services. It is based on the simple premise that consumers evaluate the value of a product/service/idea (real or hypothetical) by combining the separate amounts of value provided by each attribute. Utility, which is the conceptual basis for measuring value in conjoint analysis, is a subjective judgment of preference unique to each individual. It encompasses all product or service features, both tangible and intangible, and as such is a measure of overall preference. In conjoint analysis, utility is assumed to be based on the value placed on each of the levels of the attributes and expressed in a relationship reflecting the manner in which the utility is formulated for any combination of attributes. For example, we might sum the utility values associated with each feature of a product or service to arrive at an overall utility. Then we would assume that products or services with higher utility values are more preferred and have a better chance of choice.

Unique Aspects of Conjoint Analysis

Conjoint analysis is unique among multivariate methods in that the experimenter first constructs a set of real or hypothetical products or services by combining selected levels of each attribute. These combinations are then presented to respondents, who provide only their overall evaluations. Thus, the experimenter is asking the respondent to perform a very realistic task — choosing among a set of products. Respondents need not tell the experimenter anything else, such as how important an individual attribute is to them or how well the product performs on any specific attribute. Because the experimenter constructed the hypothetical products or services in a specific manner, the influence of each attribute and each value of each attribute on the utility judgment of a respondent can be determined from the respondents' overall ratings.

To be successful, the researcher must be able to describe the product or service in terms of both its attributes and all relevant values for each attribute. We use the term factor when describing a specific attribute or other characteristic of the product or service. The possible values for each factor are called levels. In conjoint terms, we describe a product or service in terms of its level on the set of factors characterizing it. For example, brand name and price might be two factors in a conjoint analysis. Brand name might have two levels (brand X and brand Y), whereas price might have four levels (39, 49, 59, and 69 cents). When the researcher selects the factors and the levels to describe a product or service according to a specific plan, the combination is known as a treatment or stimulus. Therefore, a stimulus for our simple example might be brand X at 49 cents.

Uses of Conjoint Analysis

The flexibility of conjoint analysis gives rise to its application in almost any area in which decisions are studied. Conjoint analysis assumes that any set of objects (e.g., brands, companies) or concepts (e.g., positioning, benefits, images) is evaluated as a bundle of attributes. Having determined the contribution of each factor to the consumer's overall evaluation, the marketing researcher could then:

1. Define the object or concept with the optimum combination of features
2. Show the relative contributions of each attribute and each level to the overall evaluation of the object
3. Use estimates of purchaser or customer judgments to predict preferences among objects with differing sets of features (other things held constant)
4. Isolate groups of potential customers who place differing importance on the features to define high and low potential segments
5. Identify marketing opportunities by exploring the market potential for feature combinations not currently available

The knowledge of the preference structure for each individual allows the researcher almost unlimited flexibility in examining both individual and aggregate reactions to a wide range of product- or service-related issues.

WHAT IS CANONICAL CORRELATION?

Whereas multiple regression analysis can predict the value of a single (metric) dependent variable from a linear function of a set of independent variables, for some research problems, interest may not center on a single dependent variable; rather, the experimenter may be interested in relationships between sets of multiple dependent and multiple independent variables. Canonical correlation analysis is a multivariate statistical model that facilitates the study of interrelationships among sets of multiple dependent variables and multiple independent variables (Green, 1978; Green and Carroll, 1978). Whereas multiple regression predicts a single dependent variable from a set of multiple independent variables, canonical correlation simultaneously predicts multiple dependent variables from multiple independent variables.

Canonical correlation places the fewest restrictions on the types of data on which it operates. Because the other techniques impose more rigid restrictions, it is generally believed that the information obtained from them is of higher quality and may be presented in a more interpretable manner. For this reason, many researchers view canonical correlation as a last-ditch effort, to be used when all other higher-level techniques have been exhausted. But in situations with multiple dependent and independent variables, canonical correlation is the most appropriate and powerful multivariate technique. It has gained acceptance in many fields and represents a useful tool for multivariate analysis, particularly as interest has spread to considering multiple dependent variables.

WHAT IS CLUSTER ANALYSIS?

Cluster analysis is the name for a group of multivariate techniques whose primary purpose is to group objects based on the characteristics they possess. Cluster analysis classifies objects (e.g., respondents, products, or other entities) so that each object is very similar to others in the cluster with respect to some predetermined selection criterion. The resulting clusters of objects should then exhibit high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity. Thus, if the classification is successful, the objects within clusters will be close together when plotted geometrically, and different clusters will be far apart.

In cluster analysis, the concept of the variate is again a central issue, but in a quite different way from other multivariate techniques. The cluster variate is the set of variables representing the characteristics used to compare objects in the cluster analysis. Because the cluster variate includes only the variables used to compare objects, it determines the "character" of the objects. Cluster analysis is the only multivariate technique that does not estimate the variate empirically but instead uses the variate as specified by the experimenter. The focus of cluster analysis is on the comparison of objects based on the variate, not on the estimation of the variate itself. This makes the experimenter's definition of the variate a critical step in cluster analysis.

Cluster analysis has been referred to as Q analysis, typology construction, classification analysis, and numerical taxonomy. This variety of names is due in part to the usage of clustering methods in such diverse disciplines as psychology, biology, sociology, economics, engineering, and business. Although the names differ across disciplines, the methods all have a common dimension: classification according to natural relationships (Aldenderfer and Blashfield, 1984; Anderburg, 1973; Bailey, 1994; Sneath and Sokal, 1973; Everitt, 1980). This common dimension represents the essence of all clustering approaches. As such, the primary value of cluster analysis lies in the classification of data, as suggested by "natural" groupings of the data themselves. Cluster analysis is comparable to factor analysis in its objective of assessing structure. But cluster analysis differs from factor analysis in that cluster analysis groups objects, whereas factor analysis is primarily concerned with grouping variables.

Cluster analysis is a useful data analysis tool in many different situations. For example, a researcher who has collected data by means of a questionnaire may be faced with a large number of observations that are meaningless unless classified into manageable groups. Cluster analysis can perform this data reduction procedure objectively by reducing the information from an entire population or sample to information about specific, smaller subgroups. For example, if we can understand the attitudes of a population by identifying the major groups within the population, then we have reduced the data for the entire population into profiles of a number of groups. In this fashion, the researcher has a more concise, understandable description of the observations, with minimal loss of information.

Cluster analysis is also useful when a researcher wishes to develop hypotheses concerning the nature of the data or to examine previously stated hypotheses. For example, an engineer may believe that attitudes toward performance of a car versus comfortable ride could be used to separate consumers into logical segments or

groups. Cluster analysis can classify the performance consumers by their attitudes versus consumers who prefer comfort, and the resulting clusters, if any, can be profiled for demographic similarities and differences.

These examples are just a small fraction of the types of applications of cluster analysis. Ranging from the derivation of taxonomies in biology for grouping all living organisms, to psychological classifications based on personality and other personal traits, to segmentation analyses of marketers, cluster analysis has always had a strong tradition of grouping people. This tradition has been extended to classifying objects, including the market structure, analyses of the similarities and differences among new products, and performance evaluations of firms to identify groupings based on the firms' strategies or strategic orientations. The result has been an explosion of applications in almost every area of inquiry, creating not only a wealth of knowledge on the use of cluster analysis but also the need for a better understanding of the technique to minimize its misuse.

Yet, along with the benefits of cluster analysis come some caveats. Cluster analysis can be characterized as descriptive, atheoretical, and noninferential. Cluster analysis has no statistical basis upon which to draw statistical inferences from a sample to a population, and it is used primarily as an exploratory technique. The solutions are not unique, as the cluster membership for any number of solutions is dependent upon many elements of the procedure, and many different solutions can be obtained by varying one or more elements. Moreover, cluster analysis will always create clusters, regardless of the "true" existence of any structure in the data. Finally, the cluster solution is totally dependent upon the variables used as the basis for the similarity measure. The addition or deletion of relevant variables can have a substantial impact on the resulting solution. Thus, the experimenter must take particular care in assessing the impact of each decision involved in performing a cluster analysis.

WHAT IS MULTIDIMENSIONAL SCALING?

Multidimensional scaling (MDS), also known as perceptual mapping, is a procedure that allows an experimenter to determine the perceived relative image of a set of objects (firms, products, ideas, or other items associated with commonly held perceptions). The purpose of MDS is to transform consumer judgments of similarity or preference (e.g., preference for stores or brands) into distances represented in multidimensional space. Assume that objects A and B are judged by respondents to be the most similar compared with all other possible pairs of objects. MDS techniques will position objects A and B so that the distance between them in multidimensional space is smaller than the distance between any other two pairs of objects. The resulting perceptual map, also known as a spatial map, shows the relative positioning of all objects.

Multidimensional scaling is based on the comparison of objects. Any object (e.g., product, service, image, aroma) can be thought of as having both perceived and objective dimensions. For example, Thisvi's management may see their product (a car) as having two color options (red and green), a 100-horsepower motor, and a 124 inches wheel-to-wheel base. These are the objective dimensions. On the other hand, customers may (or may not) see these attributes. Customers may also perceive

the Thisvi car as expensive-looking or fragile. These are perceived dimensions, also known as subjective dimensions. Two products may have the same physical characteristics (objective dimensions) but be viewed differently because the different brands are perceived to differ in quality (a perceived dimension) by many customers. Thus, the following two differences between objective and perceptual dimensions are very important:

1. *Individual Differences:* The dimensions perceived by customers may not coincide with (or may not even include) the objective dimensions assumed by the experimenter. We expect that each individual may have different perceived dimensions, but the experimenter must also accept that the objective dimensions may also vary substantially. Individuals may consider different sets of objective characteristics as well as vary the importance they attach to each dimension.
2. *Interdependence:* The evaluations of the dimensions (even if the perceived dimensions are the same as the objective dimensions) may not be independent and may not agree. Both perceived and objective dimensions may interact with one another to create unexpected evaluations. For example, one soft drink may be judged sweeter than another because the first has a fruitier aroma, although both contain the same amount of sugar.

The challenge to the experimenter is first to understand the perceived dimensions and then to relate them to objective dimensions, if possible. Additional analysis is needed to assess which attributes predict the position of each object in both perceptual and objective space.

A note of caution must be raised, however, concerning the interpretation of dimensions. Because this process is more an art than a science, the experimenter must resist the temptation to allow personal perception to affect the qualitative dimensionality of the perceived dimensions. Given the level of researcher input, caution must be taken to be as objective as possible in this critical, yet still rudimentary, area.

WHAT IS STRUCTURAL EQUATION MODELING?

Structural equation modeling (SEM) encompasses an entire family of models known by many names, among them covariance structure analysis, latent variable analysis, confirmatory factor analysis, and often simply LISREL analysis (the name of one of the more popular software packages). Resulting from an evolution of multiequation modeling developed principally in econometrics and merged with the principles of measurement from psychology and sociology, SEM has emerged as an integral tool in both managerial and academic research (Austin and Calderon, 1996; Bagozzi and Yi, 1988; Bentler, 1980; Breckler, 1990; Dolan, 1996; Duncan, 1975; Fan, 1997; Hatcher, 1996; Hox, 1995; Joreskog and Sorbom, 1993a; Marsh and Hoceuar, 1994; McDonald and Marsh, 1990; Neale et al., 1989; O'Brien and Reilly, 1995; Predhazur and Schmelkin, 1992; Rigton, 1996; Robles, 1996; Rubio and Gillespie, 1995; Steenkamp and van Trijp, 1991; Tremblay and Gardner, 1996) can

also be used as a means of estimating other multivariate models, including regression, principal components (Dolan, 1996), canonical correlation (Fan, 1997), and even MANOVA (Bagozzi, 1988).

As might be expected for a technique with such widespread use and so many variations in applications, many researchers are uncertain about what constitutes structural equation modeling. Yet all SEM techniques are distinguished by two characteristics: estimation of multiple and interrelated dependence relationships, and the ability to represent unobserved concepts in these relationships and account for measurement error in the estimation process.

Accommodating Multiple Interrelated Dependence Relationships

The most obvious difference between SEM and other multivariate techniques is the use of separate relationships for each of a set of dependent variables. In simple terms, SEM estimates a series of separate, but interdependent, multiple regression equations simultaneously by specifying the structural model used by the statistical program. First, the experimenter draws upon theory, prior experience, and the research objectives to distinguish which independent variables predict each dependent variable. For example, we may first want to predict "car" image. We then may want to use "car" image to predict satisfaction, both of which in turn may be used to predict "car" loyalty. Thus, some dependent variables become independent variables in subsequent relationships, giving rise to the interdependent nature of the structural model. Moreover, many of the same variables affect each of the dependent variables, but with differing effects. The structural model expresses these relationships among independent and dependent variables, even when a dependent variable becomes an independent variable in other relationships.

The proposed relationships are then translated into a series of structural equations (similar to regression equations) for each dependent variable. This feature sets SEM apart from techniques discussed previously that accommodate multiple dependent variables — multivariate analysis of variance and canonical correlation — in that they allow only a *single* relationship between dependent and independent variables.

Incorporating Variables that We Do Not Measure Directly

The estimation of multiple interrelated dependence relationships is not the only unique element of structural equation modeling. SEM also has the ability to incorporate latent variables into the analysis. A latent variable is a hypothesized and unobserved concept that can only be approximated by observable or measurable variables. The observed variables, which we gather from respondents through various data collection methods (e.g., surveys, tests, observations), are known as manifest variables. Yet why would we want to use a latent variable that we did not measure instead of the exact data (manifest variables) the respondents provided? Although this may sound like a nonsensical or "black box" approach, it has both practical and theoretical justification by improving statistical estimation, better representing theoretical concepts, and accounting for measurement error.

Improving Statistical Estimation

Statistical theory tells us that a regression coefficient is actually composed of two elements: the "true" or structural coefficient between the dependent and independent variable and the reliability of the predictor variable. Reliability is the degree to which the independent variable is "error-free" (Blalock, 1982). In all the multivariate techniques to this point, we have assumed we had no error in our variables. But we know from both practical and theoretical perspectives that we cannot perfectly measure a concept and that there is always some degree of measurement error. For example, when asking about something as straightforward as household income, we know some people will answer incorrectly, either overstating or understating the amount or not knowing it precisely. The answers provided have some measurement error and thus affect the estimation of the "true" structural coefficient (Rigdon, 1994).

The impact of measurement error (and the corresponding lowered reliability) can be shown from an expression of the regression coefficient as

$$\beta_{yx} = \beta_s X \rho_x$$

where β_{yx} is the observed regression coefficient, β_s is the "true" structural coefficient, and ρ_x is the reliability of the predictor variable. Unless the reliability is 100%, the observed correlation will always understate the "true" relationship. Because all dependence relationships are based on the observed correlation (and resulting regression coefficient) between variables, we would hope to "strengthen" the correlations used in the dependence models and make them more accurate estimates of the structural coefficients by first accounting for the correlation attributable to any number of measurement problems.

Overall Goodness-of-Fit Measures for Structural Equation Modeling

Assessing the overall goodness-of-fit for structural equation models is not as straightforward as with other multivariate dependence techniques such as multiple regression, discriminant analysis, multivariate analysis of variance, or even conjoint analysis. SEM has no single statistical test that best describes the "strength" of the model's predictions. Instead, experimenters have developed a number of goodness-of-fit measures that, when used in combination, assess the results from three perspectives: overall fit, comparative fit to a base model, and model parsimony. The discussions that follow present alternative measures for each of these perspectives, along with the methods of calculation for those measures that are not contained in the results and that must be computed separately.

One common question arises in the discussion of each measure: What is an acceptable level of fit? None of the measures (except the chi-square statistic) has an associated statistical test. Although in many instances guidelines have been suggested, no absolute test is available, and the experimenter must ultimately decide whether the fit is acceptable. Bollen (1989, p. 275) addresses this issue directly: "Overall, selecting a rigid cutoff for the incremental fit indices is like selecting a

minimum R^2 for a regression equation. Any value will be controversial. Awareness of the factors affecting the values and good judgment are the best guides to evaluating their size." This advice applies equally well to the other goodness-of-fit measures.

Before examining the various goodness-of-fit measures, it may be useful to review the derivation of degrees of freedom in structural models. The number of unique data values in the input matrix is s (where $s = {}^1/_2(k)(k-1)$ and k is the total number of indicators for both endogenous and exogenous constructs). The degrees of freedom (df) for any estimated model are then calculated as $df = s - t$, where t is the number of estimated coefficients. If the experimenter knows the df for an estimated model and the total number of indicators, then t can be calculated directly as $t = s - df$.

The examination and derivation of goodness-of-fit measures for SEM has gained widespread interest among academic researchers in recent years, resulting in the continual development of new goodness-of-fit measures (Ding et al., 1995; Rigdon, 1994, 1996; Tanaka, 1993; Satorra and Bentler, 1994; and others). This is reflected in the statistical programs as they are continually modified to provide the most relevant information regarding the estimated model. In this discussion, we have focused our attention on the LISREL program because of its widespread application. It has undergone these changes as well. The newest version of LISREL substantially expands the number and type of fit indices available directly in the output. For this reason, the following discussion and example data detail the calculations of those measures not provided in earlier versions of the program.

Measures of Absolute Fit

Absolute fit measures determine the degree to which the overall model (structural and measurement models) predicts the observed covariance or correlation matrix. No distinction is made as to whether the model fit is better or worse in the structural or measurement models. Among the absolute fit measures commonly used to evaluate SEM are the chi-square statistic, the noncentrality parameter, the goodness-of-fit statistic, the root mean square error, the root mean square error of approximation, and the expected cross-validation index.

Likelihood-Ratio Chi-Square Statistic

The most fundamental measure of overall fit is the likelihood-ratio chi-square (χ^2) statistic, the only statistically based measure of goodness-of-fit available in SEM (Joreskog and Sorbom, 1993b). A large value of chi-square relative to the degrees of freedom signifies that the observed and estimated matrices differ considerably. Statistical significance levels indicate the probability that these differences are due solely to sampling variations. Thus, low chi-square values, which result in significance levels greater than .05 or .01, indicate that the actual and predicted input matrices are not statistically different. In this instance, the experimenter is looking for *nonsignificant differences* because the test is between actual and predicted matrices. The experimenter must remember that this method differs from the customary desire to find statistical significance. However, even statistical nonsignificance does

not guarantee that the "correct" model has been identified, but only that this proposed model fits the observed covariances and correlations well. It does not assure the experimenter that another model would not fit as well or better. The .05 significance level is recommended as the minimum accepted, and levels of .1 or .2 should be exceeded before nonsignificance is confirmed (Fornell, 1983).

An important criticism of the chi-square measure is that it is too sensitive to sample size differences, especially for cases in which the sample size exceeds 200 respondents. As sample size increases, this measure has a greater tendency to indicate significant differences for equivalent models. If the sample size becomes large enough, significant differences will be found for any specified model. Moreover, as the sample size nears 100 or goes even lower, the chi-square test will show acceptable fit (nonsignificant differences in the predicted and observed input matrices), even when none of the model relationships is shown to be statistically significant. Thus, the chi-square statistic is quite sensitive in different ways to both small and large sample sizes, and the experimenter is encouraged to complement this measure with other measures of fit in all instances. The use of chi-square is appropriate for sample sizes between 100 and 200, with the significance test becoming less reliable with sample sizes outside this range.

The sensitivity of the chi-square measure extends past sample size considerations. For example, it has been shown that this measure varies based on the number of categories in the response variable (Green et al., 1997). Given its sensitivity to many factors, the researcher is encouraged to complement the chi-square measure with other goodness-of-fit measures.

Noncentrality and Scaled Noncentrality Parameters

The noncentrality parameter (NCP) is the result of statisticians' search for an alternative measure to the likelihood-ratio chi-square statistic that is less affected by or independent of the sample size. Statistical theory suggests that a noncentrality chi-square measure will be less affected by sample size in its representation of the differences between the actual and estimated data matrices (McDonald and Marsh, 1990). In a LISREL problem, the noncentrality parameter can be calculated as:

$$\text{NCP} = \chi^2 - \text{Degrees of freedom}$$

Although this measure adjusts the chi-square by the degrees of freedom of the estimated model, it is still in terms of the original sample size. To "standardize" the NCP, divide it by the sample size to obtain the scaled noncentrality parameter (SNCP) (McDonald and Marsh, 1990). This can be calculated as

$$\text{SNCP} = [\chi^2 - \text{Degrees of freedom}]/\text{Sample size}$$

This scaled measure is analogous to the average squared Euclidean distance measure between the estimated model and the unrestricted model (McDonald and Marsh, 1990). For both the unscaled and the scaled parameters, the objective is to

minimize the parameter value. Because there is no statistical test for this measure, it is best used in making comparisons between alternative models.

Goodness-of-Fit Index

The goodness-of-fit index (Joreskog and Sorbom, 1988b; Joreskog and Sorbom, 1993a) is another measure provided by LISREL. It is a nonstatistical measure ranging in value from 0 (poor fit) to 1.0 (perfect fit). It represents the overall degree of fit (the squared residuals from prediction compared with the actual data), but it is not adjusted for the degrees of freedom. Higher values indicate better fit, but no absolute threshold levels for acceptability have been established.

Root Mean Square Residual (RMSR)

The root mean square residual is the square root of the mean of the squared residuals — an average of the residuals between observed and estimated input matrices. If covariances are used, the RMSR is the average residual covariance. If a correlation matrix is used, then the RMSR is in terms of an average residual correlation. The RMSR is more useful for correlations, which are all on the same scale, than for covariances, which may differ from variable to variable depending on unit of measure. Again, no threshold level can be established, but the experimenter can assess the practical significance of the magnitude of the RMSR in light of the research objectives and the observed or actual covariances or correlations (Bagozzi and Yi, 1988).

Root Mean Square Error of Approximation

Another measure that attempts to correct for the tendency of the chi-square statistic to reject any specified model with a sufficiently large sample is the root mean square error of approximation (RMSEA). Similar to the RMSR, the RMSEA is the discrepancy per degree of freedom. It differs from the RMSR, however, in that the discrepancy is measured in terms of the population, not just the sample used for estimation (Steiger, 1990). The value is representative of the goodness-of-fit that could be expected if the model were estimated in the population, not just the sample drawn for the estimation. Values ranging from .05 to .08 are deemed acceptable. An empirical examination of several measures found that the RMSEA was best suited to use in a confirmatory or competing models strategy with larger samples (Rigdon, 1996).

Expected Cross-Validation Index

The expected cross-validation index (ECVI) is an approximation of the goodness-of-fit the estimated model would achieve in another sample of the same size. Based on the sample covariance matrix, it takes into account the actual sample size and the difference that could be expected in another sample. The ECVI also takes into account the number of estimated parameters for both the structural and measurement models. The ECVI is calculated as

ECVI = χ^2/Sample size – 1 + (2 × number of estimated parameters)/Sample size – 1

The EVCI has no specified range of acceptable values, but it is used in comparisons between alternative models.

Cross-Validation Index

The cross-validation index (CVI) assesses goodness-of-fit when an actual cross-validation has been performed. Cross-validation is performed in two steps. First, the overall sample is split into two samples — an estimation sample and a validation sample. The estimation sample is used to estimate a model and create the estimated correlation of covariance matrix. This matrix is then compared to the sample from the validation sample. A double cross-validation process can be performed by comparing the estimated correlation or covariance matrix from each sample to a data matrix from the other sample.

Incremental Fit Measures

The second class of measures compares the proposed model to some baseline model, most often referred to as the null model. The null model should be some realistic model that all other models should be expected to exceed. In most cases, the null model is a single-construct model with all indicators perfectly measuring the construct (i.e., this represents the chi-square value associated with the total variance in the set of correlations or covariances). There is, however, some disagreement over exactly how to specify the null model in many situations (Sobel and Bohrnstedt, 1985).

Adjusted Goodness-of-Fit Index

The adjusted goodness-of-fit is an extension of the GFI, adjusted by the ratio of degrees of freedom for the proposed model to the degrees of freedom for the null model. It is quite similar to the parsimonious normed fit index and a recommended acceptance level is a value greater than or equal to .90.

Tucker-Lewis Index

The next incremental fit measure is the Tucker-Lewis index (Tucker and Lewis, 1973), also known as the nonnormed fit index (NNFI). First proposed as a means of evaluating factor analysis, the TLI has been extended to SEM. It combines a measure of parsimony into a comparative index between the proposed and null models, resulting in values ranging from 0 to 1.0. It is expressed as:

$$\text{TLI} = [(\chi^2_{null} / df_{null}) - (\chi^2_{proposed} / df_{proposed})](\chi^2_{null} / df_{null}) - 1$$

A recommended value of TLI is .90 or greater. This measure can also be used for comparing between alternative models by substituting the alternative model for the null model.

Normed Fit Index

One of the more popular measures is the normed fit index (Bentler and Bonnett, 1980), which is a measure ranging from 0 (no fit at all) to 1.0 (perfect fit). Again, the NFI is a relative comparison of the proposed model to the null model. The NFI is calculated as:

$$\text{NFI} = \frac{\chi^2_{null} - \chi^2_{proposed}}{\chi^2_{null}}$$

As with the Tucker-Lewis index, there is no absolute value indicating an acceptable level of fit, but a commonly recommended value is .90 or greater.

Other Incremental Fit Measures

A number of other incremental fit measures have been proposed, and the newer version of LISREL includes three in its output. The relative fit index (RFI), the incremental fit index (IFI), and the comparative fit index (CFI) all represent comparisons between the estimated model and a null or independence model. The values lie between 0 and 1.0, and larger values indicate higher levels of goodness-of-fit. The CFI has been found to be more appropriate in a model development strategy or when a smaller sample is available (Rigdon, 1996). The interested reader can find the specific details of each measure in selected readings (Bollen, 1986 and 1989; Bentler, 1990).

Parsimonious Fit Measures

Parsimonious fit measures relate the goodness-of-fit of the model to the number of estimated coefficients required to achieve this level of fit. Their basic objective is to diagnose whether model fit has been achieved by "overfitting" the data with too many coefficients. This procedure is similar to the "adjustment" of the R^2 in multiple regression. However, because no statistical test is available for these measures, their use in an absolute sense is limited in most instances to comparisons between models.

Parsimonious Normed Fit Index

The first measure in this case is the parsimonious normed fit index (PNFI) (James et al., 1982), a modification of the NFI. The PNFI takes into account the number of degrees of freedom used to achieve a level of fit. Parsimony is defined as achieving higher degrees of fit per degree of freedom used (one degree of freedom per estimated coefficient). Thus more parsimony is desirable. The PNFI is defined as:

$$\text{PNFI} = \frac{df_{proposed}}{df_{null}} \times \text{NFI}$$

Higher values of PNFI are better, and its principal use is for the comparison of models with differing degrees of freedom. It is used to compare alternative models, and there are no recommended levels of acceptable fit. However, when comparing between models, differences of .06 to .09 are proposed to be indicative of substantial model differences (Williams and Holahan, 1994).

Parsimonious Goodness-of-Fit Index

The parsimonious goodness-of-fit index (PGFI) modifies the GFI differently from the AGFI. Where the AGFI's adjustment of the GFI was based on the degrees of freedom in the estimated and null models, the PGFI is based on the parsimony of the estimated model. It adjusts the GFI in the following manner:

$$\text{PGFI} = [df_{proposed}]/{}^{1}/_{2}\ \text{(No. of manifest variables)(No. of manifest variables + 1)} \times \text{GFI}$$

The value varies between 0 and 1.0, with higher values indicating greater model parsimony.

Normed Chi-Square

Joreskog (1970) proposed that the chi-square be "adjusted" by the degrees of freedom to assess model fit for various models. This measure can be termed the normed chi-square and is the ratio of the chi-square divided by the degrees of freedom. This measure provides two ways to assess inappropriate models: (1) a model that may be "overfitted," thereby capitalizing on chance, typified by values less than 1.0; and (2) models that are not yet truly representative of the observed data and thus need improvement, having values greater than an upper threshold, either 2.0 or 3.0 (Carmines and McIver, 1981) or the more liberal limit of 5.0 (Joreskog, 1970). However, because the chi-square value is the major component of this measure, it is subject to the sample size effects discussed earlier with regard to the chi-square statistic.

The normed chi-square has been shown to be somewhat unreliable (Hayduk, 1987; Wheaton, 1987), so experimenters should always combine it with other goodness-of-fit measures.

Akaike Information Criterion

Another measure based on statistical information theory is the Akaike information criterion (AIC; Akaike, 1987). Similar to the PNFI, the AIC is a comparative measure between models with differing numbers of constructs. The AIC is calculated as:

$$\text{AIC} = \chi^2 + 2 \times \text{Number of estimated parameters}$$

AIC values closer to zero indicate better fit and greater parsimony. A small AIC generally occurs when small chi-square values are achieved with fewer estimated coefficients. This shows not only a good fit of observed versus predicted covariances or correlations but also a model not prone to "overfitting."

REFERENCES

Akaike, H., Factor analysis and AIC, *Psychometrika*, 52, 317–332, 1987.

Aldenderfer, M.S. and Blashfield, R.K., *Cluster Analysis*, Sage Publications, Thousand Oaks, CA, 1984.

Anderburg, M., *Cluster Analysis for Applications*, Academic Press, New York, 1973.

Austin, J.T. and Calderon, R.F., Theoretical and technical contributions to structural equation modeling: an updated annotated bibliography, *Structural Equation Modeling,* 3, 105–125, 1996.

Bailey, K.D., *Typologies and Taxonomies: An Introduction to Classification Techniques*, Sage Publications, Thousand Oaks, CA, 1994.

Bentler, P.M., Multivariate analysis with latent variables: causal modeling, *Annual Review of Psychology*, 31, 419–456, 1980.

Bentler, P.M., Comparative fit indexes in structural models, *Psychological Bulletin*, 107, 238–246, 1990.

Bentler, P.M. and Bonnett, D.G., Significance tests and goodness of fit in the analysis of covariance structures, *Psychological Bulletin*, 88, 588–606, 1980.

Blalock, H.M., *Conceptualization and Measurement in the Social Sciences*, Sage, Beverly Hills, CA, 1982.

Bagozzi, R.P. and Yi, Y., On the use of structural equation models in experimental designs, *Journal of Marketing Research*, 26, 271–284, 1988.

Bollen, K.A., Sample size and Bentler and Bonnett's nonnormed fit index, *Psychometrica*, 51, 375–377, 1986.

Bollen, K.A., *Structural Equations with Latent Variables*, John Wiley & Sons, New York, 1989.

Breckler, S.J., Applications of covariance structure modeling in psychology: cause for concern? *Psychological Bulletin,* 107, 260–273, 1990.

Carmines, E. and McIver, J., Analyzing models with unobserved variables: analysis of covariance structures, in *Social Measurement: Current Issues*, Bohrnstedt, G. and Borgatta, E., Eds., Sage, Beverly Hills, CA, 1981.

Daniel, C. and Wood, F.S., *Fitting Equations to Data*, 2nd ed., Wiley-Interscience, New York, 1980.

Ding, L., Velicer, W.F., and Harlow, L.L., Effects of estimation methods, number of indicators per factor and improper solutions on structural equation modeling fit indices, *Structural Equation Modeling,* 2, 119–143, 1995.

Dolan, C., Principal component analysis using LISREL8, *Structural Equation Modeling*, 3, 307–322, 1996.

Duncan, O.D., *Introduction to Structural Equation Models,* Academic Press, New York, 1975.

Everitt, B., *Cluster Analysis*, 2nd ed., Halsted Press, New York, 1980.

Fan, X., Canonical correlation analysis and structural equation modeling: what do they have in common? *Structural Equation Modeling*, 4, 65–79, 1997.

Fisher, R.A., The use of multiple measurements in taxonomic problems, *Annals of Eugenics,* 7, 179–198, 1936.

Fornell, C., Issues in application of covariance structure analysis: a comment, *Journal of Consumer Research*, 9, 443–448, 1983.

Green, S.B., Akay, T.M., Fleming, K.K., Hershberger, S.C., and Marquis, J.G., Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis, *Structural Equation Modeling*, 4, 108–120, 1997.

Green, P.E., *Analyzing Multivariate Data*, Holt, Rinehart and Winston, Hinsdale, IL, 1978.

Green, P.E. and Carroll, J.D., *Mathematical Tools for Applied Multivariate Analysis*, Academic Press, New York, 1978.

Hatcher, L., Using SAS PROC CALIS for path analysis: an introduction, *Structural Equation Modeling,* 3, 176–192, 1996.

Hayduk, L.A., *Structural Equation Modeling with LISREL: Essentials and Advances,* John Hopkins University Press, Baltimore, 1987.

Hox, J.J., AMOS, EQS, and LISREL for Windows: a comparative review, *Structural Equation Modeling,* 2, 79–91, 1995.

Huberty, C.J., Discriminant analysis, *Review of Educational Research,* 45, 543–598, 1975.

Huberty, C.J., Multivariate indices of strength of association, *Multivariate Behavior Research,* 7, 523–526, 1972.

James, L.R., Muliak, S.A., and Brett, J.M., *Causal Analysis: Assumptions, Models and Data*, Sage, Beverly Hills, CA, 1982.

Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, Prentice Hall, Upper Saddle River, NJ, 1982.

Joreskog, K.G., A general method for analysis of covariance structures, *Biometrika*, 57, 239–251, 1970.

Joreskog, K.C. and Sorborn, D., *LISREL VII: Analysis of Linear Structure Relationships by the Method of Maximum Likelihood,* Scientific Software, Mooresville, IL, 1988.

Joreskog, K.G. and Sorbom D., *PRELIS2: A Program for Multivariate Data Screening and Data Summarization*, Scientific Software, Mooresville, IL, 1993a.

Joreskog, K.G. and Sorbom, D., *LISREL8: Structural Equation Modeling with SIMPLIS Command Language,* Scientific Software, Mooresville, IL, 1993b.

Klecka, W.R., *Discriminant Analysis,* Sage, Beverly Hills, CA, 1980.

Lindeman, R.H., Merenda, P.F., and Gold, R.Z., *Introduction to Bivariate and Multivariate Analysis,* Scott Foresman, Glenview, IL, 1980.

Marsh, H.W. and Hoceuar, D., Confirmatory factor analysis models of factorial invariance: a multifaceted approach, *Structural Equation Modeling*, 1, 5–34, 1994.

McDonald, R.P. and Marsh, H.W., Choosing a multivariate model: noncentrality and goodness of fit, *Psychological Bulletin*, 107, 247–255, 1990.

Neale, M.C., Heath, A.C., Kewitt, J.K., Eaves, L.J., and Walker, D.W., Fitting genetic models with LISREL: hypothesis testing, *Behavior Genetics*, 19, 37–49, 1989.

O'Brien, R.M. and Reilly, T., Equality in constraints and metric-setting measurement models, *Structural Equation Modeling*, 2, 53–72, 1995.

Overall, J. E. and Klett, C.J., *Applied Multivariate Analysis,* McGraw-Hill, New York, 1972.

Predhazur, E.J. and Schmelkin, L.P., *Measurement Design and Analysis: An Integrated Approach*, Lawrence Eribaum and Associates, Hillsdale, NJ, 1992.

Rigdon, E.E., Demonstrating the effects of unmodeled random measurement error, *Structural Equation Modeling,* 1, 375–380, 1994.

Rigton, E.E., CFI versus RMSEA: A comparison of two fit indices for structural equation modeling, *Structural Equation Modeling,* 3, 369–379, 1996.

Robles, J., Confirmation bias in structural equation modeling, *Structural Equation Modeling,* 3, 307–322, 1996.

Rubio, D.M. and Gillespie, D.E., Problems with error in structural equation models, *Structural Equation Modeling*, 2, 367–378, 1995.

Rulon, P.J., Tiedeman, D.V., Tatsuoka, M.M., and Langmuir, C.R., *Multivariate Statistics for Personnel Classification,* John Wiley & Sons, New York, 1967.

Satorra, A. and Bentler, P., Correction to test statistics and standard errors in covariance structure analysis, in *Latent Variable Analysis: Applications for Development Research,* Von Eye, A. and Clogg, C., Eds., Sage, Newbury Park, CA, 1994.

Shaffer, J.P. and Gillo, M.W., A multivariate extension of the correlation ratio, *Educational and Psychological Measurement,* 34, 521–524, 1974.

Smith, I.L., The eta coefficient in MANOVA, *Multivariate Behavioral Research,* 7, 361–372, 1972.

Sneath, P.H.A. and Sokal, R.R., *Numerical Taxonomy,* Freeman Press, San Francisco, 1973.

Sobel, M.E. and Bohrnstedt, G.W., The use of null models in evaluating the fit of covariance structure models, in *Sociological Methodology*, Tuma, N.B., Ed., Jossey-Bass, San Francisco, 1985.

Steenkamp, J.E.M. and van Trijp, H.C.M., The use of LISREL in validating marketing constructs, *International Journal of Research in Marketing*, 8, 283–299, 1991.

Steiger, J.H., Structural model evaluation and modification: an interval estimation approach, *Multivariate Behavioral Research*, 25, 173–180, 1990.

Stevens, J.P., Global measures of association in multivariate analysis of variance, *Multivariate Behavioral Research,* 7, 373–378, 1972.

Tanaka, J., Multifaceted conceptions of fit in structural equation models, in *Testing Structural Equation Models*, Bollen, K.A. and Long, J.S., Eds., Sage, Newbury Park, CA, 1993.

Tatsuoka, M.M., *Discriminant Analysis*, Institute for Personality and Ability Testing, Champaign, IL, 1970.

Tatsuoka, M.M., *Multivariate Analysis: Techniques for Educational and Psychological Research,* John Wiley & Sons, NY, 1971.

Tatsuoka, M.M., *Classification Procedures: Profile Similarity,* Institute for Personality and Ability Testing, Champaign, IL, 1974.

Tatsuoka, M.M., Classification procedures, in *Introductory Multivariate Analysis,* Amick, D.J. and Walberg, H.J., Eds., McCutchan, Berkeley, CA, 1975.

Tatsuoka, M.M., Discriminant analysis, in *Data Analysis Strategies and Designs for Substance Abuse Research,* Bentler, P.M., Lettieri, D.L., and Austin, G.A., Eds., U.S. Government Printing Office, Washington, DC, 1976.

Tremblay, P.E. and Gardner, R.G., On the growth of structural equation modeling in psychological journals, *Structural Equation Modeling*, 3, 93–104, 1996.

Tucker, L.R. and Lewis, C., The reliability coefficient for maximum likelihood factor analysis, *Psychometrika*, 38, 1–10, 1973.

Van Ryzin, J., Ed., *Classification and Clustering,* Academic Press, New York, 1977.

Weisberg, S., *Applied Linear Regression,* John Wiley & Sons, New York, 1985.

Wheaton, B., Assessment of fit in overidentified models with latent variables, *Sociological Methods and Research*, 16, 118–154, 1987.

Williams, L.J. and Holahan, P.J., Parsimony-based fit indices for multiple-indicator models, *Structural Equation Modeling*, 1, 161–189, 1994.

12 Time Series and Forecasting

Forecasting is very difficult but essential in any business. Most forecasts are based on past history, and there lies one of the problems. There is no guarantee that the past will repeat itself. Another, of course, is the introduction of changes that will certainly change the outcome of the expectations. Practically all forecasts get these two principles wrong and, as a consequence, statisticians depend on the confidence levels of the forecast to save face. This chapter will introduce the reader to time series and forecasting in a cursory way. Specifically, it will cover autocorrelation, exponential trends, curve smoothing, and econometric models.

EXTRAPOLATION METHODS

Extrapolation methods are quantitative methods that use past data to forecast the future. Primarily, this form of analysis depends on the identification of patterns in the data with the hope that these patterns will be able to be projected into the future. Sometimes, these patterns are controlled by seasonal patterns or known cycles.

Although the effectiveness of extrapolation methods is not yet proven (Armstrong, 1986; Schnarrs and Bavuso, 1986), these methods are extensively used. The reason for their use is that we all want to predict what may happen in some future time given some facts — never mind that these facts are more often than not dynamic and that they change our expected results, every time.

Exponential Trend

Previously we saw that in a linear relationship we can forecast future expectations based on regression modeling. In contrast to a linear trend, an exponential trend is appropriate when the time series changes by a constant percentage (as opposed to a constant such as dollar amount) each period. Then the appropriate regression equation is

$$Y_t = ce^{bt}u_t$$

where c and b are constants and u represents a multiplicative error term. By taking logarithms of both sides, and letting $a = \ln(c)$ and $\varepsilon_t = \ln(u_t)$, we obtain a linear equation that can be estimated by the usual linear regression method. However, note that the response variable is now the logarithm of Y_t:

$$\ln(Y_t) = a + b_t + \varepsilon_t$$

Because the computer does the calculations, our main responsibility is to interpret the final result. This is not too difficult. It can be shown that the coefficient b (expressed as a percentage) is approximately the percentage change per period. For example, if b = .05, then the series is *increasing* by approximately 5% per period. On the other hand, if b = –.05, then the series is *decreasing* by approximately 5% per period. An exponential trend can be estimated with a regression procedure (simple or multiple, whichever you prefer), but only after the log transformation has been made on Y_t. Figure 12.1 shows a hypothetical sales example of a time series with (a) linear trend superimposed, (b) time series of forecast errors, and (c) time series of hypothetical sales with exponential trend superimposed.

The output shows that there is some evidence of not enough runs. The expected number of runs under randomness is 24.833, and there are only 20 runs for this series. However, the evidence is certainly not overwhelming — the p-value is only .155. If we ran this test as a one-tailed test, checking only for too few runs, then the appropriate p-value would be .078, half of the value. The conclusion in either case is that sales do not tend to "zigzag" as much as a random series would — highs tend to follow highs and lows tend to follow lows — but the evidence in favor of nonrandomness is not overwhelming.

Autocorrelation

The successive observations in a random series are probabilistically independent of one another. Many time series violate this property and are instead autocorrelated. The "auto" means that successive observations are correlated with one other. For example, in the most common form of autocorrelation, positive autocorrelation, large observations tend to follow large observations, and small observations tend to follow small observations. In this case the runs test is likely to pick it up because there will be fewer runs than expected, and the corresponding Z-value for the runs test will be significantly negative. Another way to check for the same nonrandomness property is to calculate the autocorrelations of the time series.

To understand autocorrelations it is first necessary to understand what it means to lag a time series. This concept is easy to understand in spreadsheets. Imagine you are looking at a spreadsheet for sales. To lag by 1 month, we simply "push down" the series by one row; to lag by 2 months, we push down the series by two rows; to lag by 3 months, we push down the series by three rows. Figure 12.2 shows that flow.

See column C of Figure 12.2. Note that there is a blank cell at the top of the lagged series (in cell C4). We can continue to push the series down one row at a time to obtain other lags. For example, the lag 3 version of the series appears in the range E7:E54. Now there are three missing observations at the top. Note that in

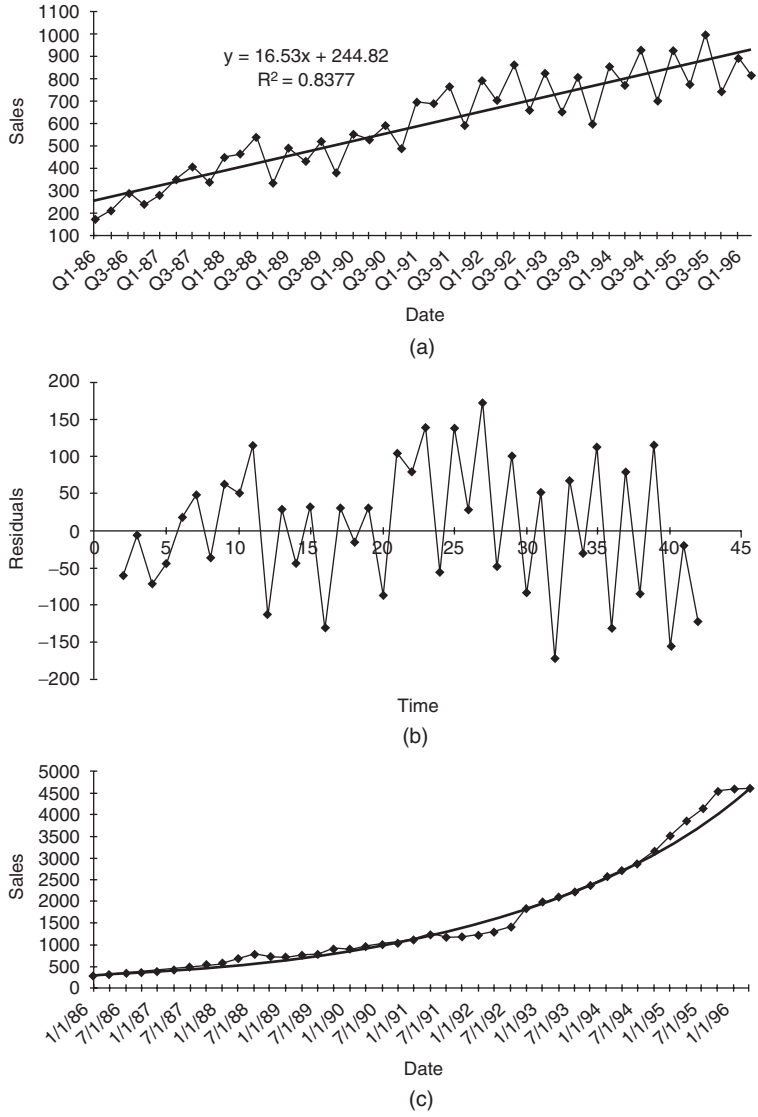


FIGURE 12.1 Time series plots.

December 1995, say, the first, second, and third lags correspond to the observations in November 1995, October 1995, and September 1995, respectively. That is, lags are simply previous observations, removed by a certain number of periods from the present time.

In general, the lag k observation corresponding to period t is Y_{t-k}. Then the autocorrelation of lag *k,* for any integer *k,* is essentially the correlation between the original series and the lag *k* version of the series. For example, in Figure 12.2, the lag 1 autocorrelation is the correlation between the observations in columns B and

	Illustration of lagging and autocorrelation									
	Month	**Sales**	**Sales_Lag1**	**Sales_Lag2**	**Sales_Lag3**		***Autocorrelations***			
	Jan-95	226	•	•	•			Lag	Autocorr	StErr
	Feb-95	254	226	•	•			1	**0.3492**	0.1443
	Mar-95	204	254	226	•			2	0.0772	0.1443
	Apr-95	193	204	254	226			3	0.0814	0.1443
	May-95	191	193	204	254			4	–0.0095	0.1443
	Jun-95	166	191	193	204			5	–0.1353	0.1443
	Jul-95	175	166	191	193			6	0.0206	0.1443
	Aug-95	217	175	166	191					
	Sep-95	167	217	175	166					
	Sep-98	175	181	179	168					
	Oct-98	185	175	181	179					
	Nov-98	245	185	175	181					
	Dec-98	177	245	185	175					

FIGURE 12.2 Lags and autocorrelation for product X (sales).

C. Similarly, the lag 2 autocorrelation is the correlation between the observations in columns B and D.

Autocorrelation may be presented in a chart format called a correlogram (see Figure 12.3).

We already mentioned that the typical autocorrelation of lag k indicates the relationship between observations k periods apart. The question, however, is how large is a "large" autocorrelation? Under the assumption of randomness, it can be shown that the standard error of any autocorrelation is approximately $1/\sqrt{T}$, in this case $1/\sqrt{48} = 0.1443$. (T = 48 which denotes the number of observations in the series.)

Exponential Smoothing

Dealing with data, many times we are willing to use moving averages to smooth our observations. When we do that, we make our forecast based on equal weight of each value in our observations. Many people would argue that if next month's forecast is to be based on the previous 12 months' observations, then more weight ought to be placed on the more recent observations. The second criticism is that the moving averages method requires a lot of data storage. This is particularly true for companies that routinely make forecasts of hundreds or even thousands of items. If 12-month moving averages are used for 1000 items, then 12,000 values are needed for next month's forecasts. This may or may not be a concern considering today's

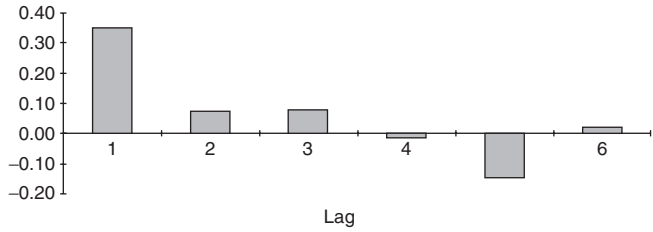


FIGURE 12.3 A typical correlogram.

relatively inexpensive computer storage capabilities. However, it is a concern of availability of data.

Exponential smoothing is a method that addresses both of these criticisms. It bases its forecasts on a weighted average of past observations, with more weight put on the more recent observations, and it requires very little data storage. In addition, it is not difficult for most business people to understand, at least conceptually. Therefore, this method finds widespread use in the business world, particularly when frequent and automatic forecasts of many items are required.

There are many versions of exponential smoothing. The simplest is called, surprisingly enough, simple exponential smoothing. It is relevant when there is no pronounced trend or seasonality in the series. If there is a trend but no seasonality, then Holt's method is applicable. If, in addition, there is seasonality, then Winters' method can be used. This does not exhaust the list of exponential smoothing models — researchers have invented many other variations — but these three models will suffice for us.

Simple Exponential Smoothing

Every exponential model has at least one smoothing constant, which is always between zero and one, and a level of the series at time t. The smoothing constant is denoted by α, and the level of the series L_t. The level value is not observable but can only be estimated. Essentially, it is where we think the series would be at time t if there were no random noise. The simple exponential smoothing method is defined by the following two equations, where F_{t+k} is the forecast of Y_{t+k} made at time t:

$$L_t = \alpha Y_t + (1-\alpha)L_{t-1}$$

$$F_{t+k} = L_t$$

Even though you usually won't have to substitute into these equations manually, you should understand what they say. The first equation shows how to update the estimate of the level. It is a weighted average of the current observation, Y_t, and the previous level, L_{t-1}, with respective weights α and $1 - \alpha$. The second equation shows how forecasts are made. It says that the k-period-ahead forecast, $F_{t+k,}$ made of Y_{t+k} in period t is the most recently estimated level, L_t.

This is the same for any value of $k \geq 1$. The idea is that in simple exponential smoothing, we believe that the series is not really going anywhere. So as soon as we estimate where the series ought to be in period t (if it weren't for random noise), we forecast that this is where it will also be in any future period.

The smoothing constant α is analogous to the span in moving averages. There are two ways to see this. The first way is to rewrite the first equation, using the fact that the forecast error, E_t, made in forecasting Y_t, at time $t - 1$ is $Y_t - F_t = Y_t - L_{t-1}$. A bit of algebra then gives

$$L_t = L_{t-1} + \alpha E_t$$

This says that the next estimate of the level is adjusted from the previous estimate by adding a multiple of the most recent forecast error. This makes sense. If our previous forecast was too high, then E_t is negative, and we adjust the estimate of the level downward. The opposite is true if our previous forecast was too low. However, this equation says that we do not adjust by the entire magnitude of E_t but only by a fraction of it. If α is small, say $\alpha = .1$, then the adjustment is minor; if α is close to 1, the adjustment is large. So if we want to react quickly to movements in the series, we choose a large α; otherwise, we choose a small α.

Another way to see the effect of α is to substitute recursively into the equation for L_t. If you are willing to go through some algebra, you can verify that L_t satisfies

$$L_t = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \alpha(1-\alpha)^3 Y_{t-3} + \ldots$$

where this sum extends back to the first observation at time t = 1. With this equation we can see how the exponentially smoothed forecast is a weighted average of previous observations. Furthermore, because $1 - \alpha$ is less than one, the weights on the Y's decrease from time t backward. Therefore, if α is close to zero, then $1 - \alpha$ is close to 1, and the weights decrease very slowly. In other words, observations from the distant past continue to have a large influence on the next forecast. This means that the graph of the forecasts will be relatively smooth, just as with a large span in the moving averages method. But when α is close to 1, the weights decrease rapidly, and only very recent observations have much influence on the next forecast. In this case, forecasts react quickly to sudden changes in the series.

What value of α should we use? There is no universally accepted answer to this question. Some practitioners recommend always using a value around .1 or .2. Others recommend experimenting with different values of alpha until you reach a measure by which the data smoothing is at optimum.

Holt's Model for Trend

The simple exponential smoothing model generally works well if there is no obvious trend in the series. But if there is a trend, then this method consistently lags behind it. For example, if the series is constantly increasing, simple exponential smoothing forecasts will be consistently low. Holt's method rectifies this by dealing with trend explicitly. In addition to the level of the series T_t, Holt's method includes a trend term, T_t and a corresponding smoothing constant β. The interpretation of T_t, is exactly as before. The interpretation of T_t is that it represents an estimate of the change in the series from one period to the next. The equations for Holt's model are as follows:

1. $L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + T_{t-1})$
2. $T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$
3. $F_{t+k} = L_t + kT_t$

These equations are not as bad as they look. (And don't forget that the computer typically does all of the calculations for you.) Equation 1 says that the updated level

is a weighted average of the current observation and the previous level plus the estimated change. Equation 2 says that the updated trend term is a weighted average of the difference between two consecutive levels and the previous trend term. Finally, equation 3 says that the k-period-ahead forecast made in period t is the estimated level plus k times the estimated change per period.

Everything we said about *a* (*alpha*) for simple exponential smoothing applies to both *a* and β (beta) in Holt's model. The new smoothing constant beta controls how quickly the method reacts to perceived changes in the trend. If beta is small, the method reacts slowly. If it is large, the method reacts more quickly. Of course, there are now two smoothing constants to select. Some practitioners suggest using a small value of *a* (.1 to .2) and setting β equal to *a.* Others suggest using an optimization option (available in most software) to select the "best" smoothing constants.

Winters' Model for Seasonality

So far we have said practically nothing about seasonality. Seasonality is defined as the consistent month-to-month (or quarter-to-quarter) differences that occur each year. For example, there is seasonality in beer sales — high in the summer months, lower in other months. Toy sales are also seasonal, with a huge peak in the months preceding Christmas. In fact, if you start thinking about time series variables that you are familiar with, the majority of them probably have some degree of seasonality.

How do we know whether there is seasonality in a time series? The easiest way is to check whether a plot of the time series has a regular pattern of ups and downs in particular months or quarters. Although random noise can sometimes obscure such a pattern, the seasonal pattern is usually fairly obvious. (Some time series software packages have special types of graphs for spotting seasonality, but we won't discuss these here.)

There are basically two extrapolation methods for dealing with seasonality. We can either use a model that takes seasonality into account explicitly and forecasts it, or we can first deseasonalize the data, then forecast the deseasonalized data, and finally adjust the forecasts for seasonality. The exponential smoothing model we discuss here, Winters' model, is of the first type. It attacks seasonality directly. Another approach is the deseasonality — the ratio to moving averages method.

Seasonal models are usually classified as additive or multiplicative. Suppose that we have monthly data, and that the average of the 12 monthly values for a typical year is 150. An additive model finds seasonal indexes, one for each month, that we add to the monthly average, 150, to get a particular month's value. For example, if the index for March is 22, then we expect a typical March value to be 150 + 22 = 172. If the seasonal index for September is –12, then we expect a typical September value to be 150 – 12 = 138. A multiplicative model also finds seasonal indexes, but we multiply the monthly average by these indexes to get a particular month's value. Now if the index for March is 1.3, we expect a typical March value to be 150(1.3) = 195. If the index for September is 0.9, then we expect a typical September value to be 150(0.9) = 135.

Either an additive or a multiplicative model can be used to forecast seasonal data. However, because multiplicative models are somewhat easier to interpret (and

have worked well in applications), we will focus on them. Note that the seasonal index in a multiplicative model can be interpreted as a percentage. Using the figures in the previous paragraph as an example, March tends to be 30% above the monthly average, whereas September tends to be 10% below it. Also, the seasonal indexes in a multiplicative model should sum to the number of seasons (12 for monthly data, 4 for quarterly data). Computer packages typically ensure that this happens.

We now turn to Winters' exponential smoothing model. It is very similar to Holt's model — it again has level and trend terms and corresponding smoothing constants alpha and beta (α, β), but it also has seasonal indexes and a corresponding smoothing constant γ (gamma). This new smoothing constant gamma controls how quickly the method reacts to perceived changes in the pattern of seasonality. If gamma is small, the method reacts slowly. If it is large, the method reacts more quickly. As with Holt's model, there are equations for updating the level and trend terms, and there is one extra equation for updating the seasonal indexes. For completeness, we list these equations below, but they are clearly too complex for hand calculation and are best left to the computer. In equation 3, S refers to the multiplicative seasonal index for period t. In equations 1, 3, and 4, M refers to the number of seasons (M = 4 for quarterly data; M = 12 for monthly data). The equations are:

1. $L_t = \alpha \dfrac{Y_t}{S_{t-M}} + (1-\alpha)(L_{t-1} + T_{t-1})$
2. $T_t - \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$
3. $S_t = \gamma \dfrac{Y_t}{L_t} + (1-\gamma)S_{t-M}$
4. $F_{t+k} = (L_t + kT_t)S_{t+k-M}$

To see how the forecasting in equation 4 works, suppose we have observed data through June and want a forecast for the coming September, that is, a 3-month-ahead forecast. (In this case t refers to June and t + k = t + 3 refers to September.) Then we first add three times the current trend term to the current level. This gives a forecast for September that would be appropriate if there were no seasonality. Next, we multiply this forecast by the most recent estimate of September's seasonal index (the one from the previous September) to get the forecast for September. Of course, the computer does all of the arithmetic, but this is basically what it is doing.

ECONOMETRIC MODELS

In the world of ever-changing demand by the customer, the pressure is on everyone to be able to forecast. Econometric models fill this need. Econometric models, also called causal models, use regression to forecast a time series variable by means of other explanatory time series variables. For example, a company might use a causal model to regress future sales on its advertising level, the population income level, the interest rate, and possibly other variables. Because we have already discussed

regression models in some depth, we will not devote much time to econometric models in this chapter; the mechanics are largely the same as in any regression analysis. However, based on the empirical evidence presented in Armstrong (1985, 1986), here are some findings:

- It is not necessary to include a lot of explanatory variables in the analysis. It is better to choose a small number of variables that, based on prior evidence, are believed to affect the response variable.
- It is important to select the proper *conceptual* explanatory variables. For example, in forecasting sales, appropriate conceptual variables might be market size, ability to buy, consumer needs, and price. However, the *operational* measures of these conceptual variables are relatively unimportant for forecast accuracy. For example, different measures of buying power might lead to comparable sales forecasts.
- Stepwise regression procedures allow the model builder to search through many possible explanatory variables to find the best model. This sounds good, and it is attractive given the access to powerful statistical software packages. However, it can lead to poor forecasts because it substitutes computer technology for sound judgment and prior theory. The moral is that we should not throw everything but the kitchen sink into the regression package and hope for the best. Some judgment regarding the variables to include may lead to better forecasts.
- High precision data on the explanatory variables are not essential. Armstrong (1986) quotes studies where models requiring forecasts of the explanatory variables did better than those where no such forecasts were required. This is contrary to intuition. We might expect that the forecasting error in the explanatory variables would lead to extra forecasting error in the response variable. This is apparently not always the case. However, the same does not hold for the response variable. It is important to have high-quality data on this variable.
- It might be a good idea to break the causal relationship into a causal chain of relationships and then estimate each part of the chain by a separate regression equation. For example, in a study of product sales, the first stage of the chain might regress price on such variables as wage rates, taxes, and warranty costs. These price predictions could then be entered into a model that predicts sales per capita as a function of personal consumption expenditures per capita and product price. The third stage could then predict the market size as a function of total population, literacy, age, and employment. Finally, the fourth stage could regress total product sales on the predicted values of product sales per capita and market size.
- The divide-and-conquer strategy outlined in the previous point is simpler than the complex simultaneous equation approach taught in many upper-level econometrics courses. The idea behind simultaneous equations is that there are several response (or endogenous) variables that cause changes in one another (Y_1 causes a change in Y_2, which in turn causes

a change in Y_1, and so on). Therefore, the regression model consists of several equations, each of which has its own response variable. According to Armstrong, despite great expenditures of time and money by the best and brightest econometricians, simultaneous equations have not been found to be of value in forecasting.

- The functional form of the relationship may not be terribly important. In particular, complex nonlinear relationships do not appear to improve forecast accuracy. However, in addition to the basic linear additive model, the constant-elasticity multiplicative model

$$Y = \alpha X_1^{b_1} X_2^{b_2} \ldots X_k^{b_k}$$

which can be transformed to a linear additive model by taking logarithms, has strong theoretical support and has done well in applications.

- A great deal of research has gone into the autocorrelation structure of econometric models. This is difficult analysis for most practitioners. Fortunately, empirical studies show that it produces only marginal gains in forecast accuracy.
- Econometric models appear to be most useful, relative to other forecasting methods, when *large* changes in the explanatory variables are expected. But in this case it is important to be able to forecast the direction of change in the explanatory variables accurately.

Armstrong summarizes his empirical findings on econometric models succinctly: "There are two important rules in the use of econometric methods: (1) keep it simple, and (2) don't make mistakes. If you obey rule 1, rule 2 becomes easier to follow." He acknowledges that rule 1 runs contrary to the thinking of many academic researchers, even practitioners. But the empirical evidence simply does not support the claim that complexity and forecast accuracy are inevitably related.

Now that we have an understanding of some of the limitations, let us explore the econometric models a little deeper. Most of the models in this chapter use only previous values of a time series variable Y to forecast future values of Y. A natural extension is to use one or more other time series variables, via a regression equation, to forecast Y. For example, if a company wants to forecast its monthly sales, it might use its own past advertising levels and/or macroeconomic variables such as Gross Domestic Product (GDP) and the prime interest rate to forecast future sales. We will not study this approach in any detail because it can become quite complex mathematically. However, there are a few points worth making.

Let X be a potential explanatory variable, such as the company's advertising level or GDP. Then either X or any of its lags could be used as explanatory variables in a regression equation for Y. If X itself is included, it is called a coincident indicator of Y_t. If only its lags are included, it is called a leading indicator of Y_t. As an example, suppose that Y_t represents a company's sales during month t and X represents its advertising level during month t. It is certainly plausible that current sales are determined more by past months' advertising levels than by the current month's

level. In this case, advertising is a leading indicator of sales, and the advertising terms that should be included in the equation are X_{t-1}, X_{t-2}, and so on. The number of lags that should be included is difficult to specify ahead of time. In practice, we might begin by including a fairly large number of lags and then discard those with insignificant coefficients in the regression output.

If we decide to use a coincident indicator, the problem is a practical one — namely, that the value of X is probably not known at the time we are forecasting Y_t. Hence, it must also be forecasted. For example, if we believe that GDP is a coincident indicator of a company's sales, then we must first forecast GDP before we can use it to forecast sales. This may be a more difficult problem than forecasting sales itself. So from a practical point of view, we would like the variables on the right-hand side of the regression equation — the explanatory variables — to be known at the time the forecast is being made.

Once we have decided which variables to use as explanatory variables, the analysis itself is carried out exactly as with any other regression analysis, and the diagnostic tools are largely the same as with regression of cross-sectional data. However, there are several things to be aware of.

First, because the explanatory variables are often lagged variables, either of the dependent variables or of some other explanatory variables, we will have to create these lagged variables to use them in the regression equation.

Second, autocorrelation may present problems. Recall that in the least squares estimation procedure, the residuals automatically average to zero. However, autocorrelation of the residuals means that errors in one period are not independent of errors in previous periods. The most common type of residual autocorrelation, *positive* autocorrelation, implies that if the forecast is on the high side in one period, it is likely to be on the high side the next period. Or if the forecast is on the low side in one period, it is likely to be on the low side the next period. We can detect residual autocorrelation by capturing the residuals and then looking at their autocorrelations to see if any are statistically significant.

Many regression outputs also include the Durbin-Watson statistic to check for lag 1 autocorrelation. The value of this statistic is always between zero and four. If the Durbin-Watson statistic is near two, then autocorrelation is not a problem. However, if it is significantly less than two (as measured by special tables), then there is significant positive autocorrelation, whereas if it is significantly greater than two, there is significant negative autocorrelation. Actually, the Durbin-Watson statistic tests only for autocorrelation of lag 1, not for any higher lags. Therefore, a graph of the autocorrelations (a correlogram) provides more complete information than is contained in the Durbin-Watson statistic alone.

If a regression model does contain significant residual autocorrelation, this is generally a sign that the model is not as good as it could be. Perhaps we have not included the best set of explanatory variables, or perhaps we have not used the best form of the response variable. This is a very complex topic, and we cannot do it justice here. Suffice it to say that a primary objective of any forecasting model, including econometric models, is to end up with uncorrelated residuals. This is easy to say, but it is often difficult to obtain, and we admit that the work necessary to

eliminate residual autocorrelation is not always worth the extra effort in terms of more accurate forecasts.

An important final point is that two time series, Y_t and X_t, may be highly correlated and hence produce a promising regression output, even though they are not really related at all. Suppose, for example, that both series are dominated by upward trends through time but are in no way related. Because they are both trending upward, it is very possible that they will have a large positive correlation, which means that the regression of Y_t on X_t has a large R^2 value. This is called a spurious correlation because it suggests a relationship that does not really exist. In such a case it is sometimes better to use an extrapolation model to model Y_t and then regress the residuals of Y_t on X_t. Intuitively, we first see how much of the behavior of the Y_t series can be explained by its own past values. Then we regress whatever remains on the X_t series.

We will not pursue this strategy here, but it does suggest how complex a rigorous econometric analysis can be. It is not just a matter of loading Ys and Xs into a computer package and running the "obvious" regression. The output could look good, but it could also be very misleading.

A FINAL COMMENT ON COMBINING FORECASTS

There is one other general forecasting method that is worth mentioning. In fact, it has attracted a lot of attention in recent years, and many researchers believe that it has great potential for increasing forecast accuracy. The method is simple — combine two or more forecasts to obtain the final forecast. The reasoning behind this method is also simple — the forecast errors from different forecasting methods may cancel one another. The forecasts that are combined can be of the same general type — extrapolation forecasts, for example — or they can be of different types, such as judgmental and extrapolation. The number of forecasts to combine and the weights to use in combining them have been the subjects of several research studies.

Although the findings are not entirely consistent, it appears that the marginal benefit from each individual forecast after the first two or three is minor. Also, there is not much evidence to suggest that the simplest weighting scheme — weight each forecast equally, that is, average them — is any less accurate than more complex weighting schemes.

REFERENCES

Armstrong, S., *Long Range Forecasting*, John Wiley & Sons, New York, 1985.

Armstrong, S., Research on forecasting: a quarter century review, 1960–1984, *Interfaces,* 16, 89–103, 1986.

Schnarrs, S. and Bavuso, J., Extrapolation models on very short term forecasts, *Journal of Business Research*, 14, 27–36, 1986.

Part II

Essential Concepts of Probability

13 Functions of Real and Random Variables

This chapter summarizes the basic concepts of real and random variables as encountered in the practical usage of statistics and probability. The attempt here is not to explain the concepts fully but rather to sensitize the reader to their significance and their application.

DETERMINISTIC MATHEMATICS (REPLICATED BY MEAN)

Engineers use deterministic mathematical formulas involving functions of variables (sums, products, powers, etc.) to describe such things as:

Geometry of objects: e.g., area, volume, arc lengths
Physical laws: e.g., F = Ma, V = iR, etc.
Processes

STATISTICAL MATHEMATICS

Generally the terms in these formulas are random variables because of uncertainties of their measurement or manufacture. We need to describe the statistical characteristics of these formulas by determining relationships between their various means and variances.

SUM OR DIFFERENCE OF TWO *REAL* VARIABLES: X_1 AND X_2

$$Y = a_1X_1 \pm a_2X_2$$

SUM OR DIFFERENCE OF TWO *RANDOM* VARIABLES: X_1 AND X_2

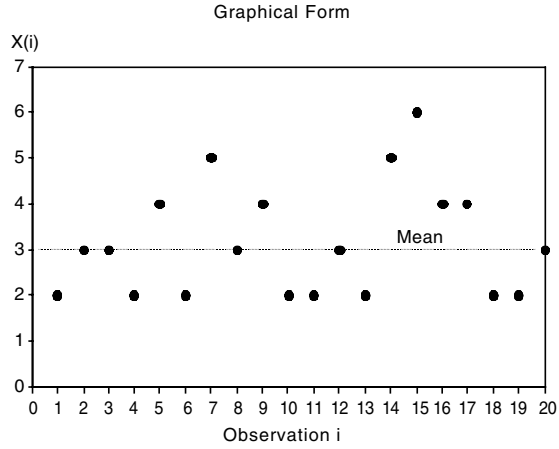
$$Y = a_1X_1 \pm a_2X_2$$

Mean: $\mu_Y = a_1\mu_{X1} \pm a_2\mu_{X2}$
Variances: $\sigma^2_Y = a_1^2\sigma^2_{X1} + a_2^2\sigma^2_{X2} \pm a_1a_2\sigma^2_{X1X2}$
Where Covariance: $\sigma^2_{X1X2} = 0$, if independent

EXAMPLE

Problem: Number of people in car pool
Experiment: Observe 20 consecutive cars in "HOV" lane
Assumption: Population is infinite
Find: Central tendencies and dispersions

Observation	Number of People X(i)
i = 1	2 = X(1)
2	3 = X(2)
3	3
4	2
5	4
6	2
7	5
8	3
9	4
10	2
11	2
12	3
13	2
14	5
15	6
16	4
17	4
18	2
19	2
i = 20	3 = $\overline{X}$(20)
n = 20	SUM = 63



Mean of observed data: (use sub i for individual sample)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X(i) = \frac{1}{20}\sum_{i=1}^{20} X(i) = \frac{1}{20}63 = 3.15 \text{ people}$$

You should not expect the mean to equal an observable value, e.g., 3.

RANK AND STACK OBSERVED DATA

1. Arrange data in ascending values.
   Assumes: order of observation not important (unlike reliability where order or time to failure is important)

$$X = \{2,2,2,2,2,2,2,2,2,\ 3,3,3,3,3,\ 4,4,4,4,\ 5,5,\ 6\}$$

2. Group ascending data in like intervals or cells

$$X = \{[2,2,2,2,2,2,2,2,2],\ [3,3,3,3,3],\ [4,4,4,4],\ [5,5],\ [6]\}$$

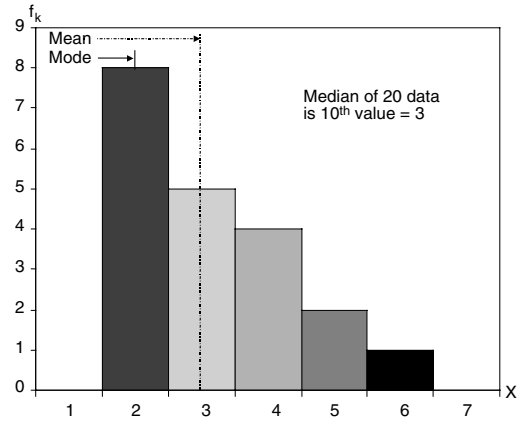
3. Assign a random variable to each cell X_k (use sub k)

$$X = \{[X_1],\ [X_2],\ [X_3],\ [X_4],\ [X_5]\}$$

4. Count "frequency" of observations in each cell, f_k

Cell X_k	Frequency f_k
$X_1 = 2$	$8 = f_1$
$X_2 = 3$	$5 = f_2$
$X_3 = 4$	$4 = f_3$
$X_4 = 5$	$2 = f_4$
$X_5 = 6$	$1 = f_5$

Mode: 2 people
Median: 3 people
Mid-Range: 4 people
Mean: $\overline{X}$ = 3.15 people

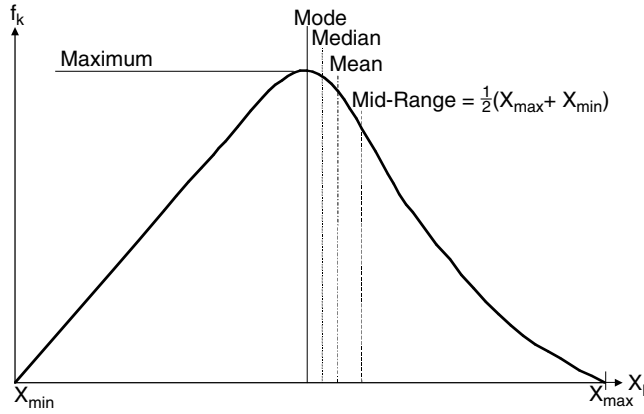


OTHER MEASURES OF CENTRAL TENDENCIES

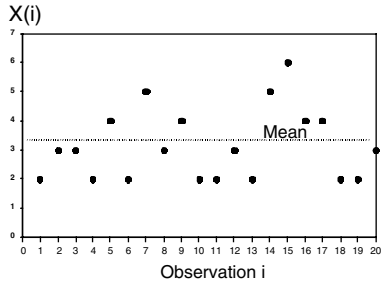
1. Mean: Average value f_k calculated from all observations (center of gravity).
   Example – Found mean X to be 3.15 people
2. Mode: Most frequently observed value of X_k, highest f_k.
   Example – Maximum f_k = 8 cars occurs when X_k = 2 people
3. Median: Value of ordered X_k that divides data in half.
   Example – Half (n/2 = 10) data points lie below, half above 3

4. Mid-Range: Value halfway between minimum and maximum observed values.

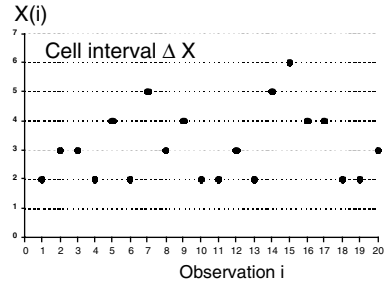
   $^1/_2$ (Maximum + Minimum) = $^1/_2$ (6 + 2) = 4 people



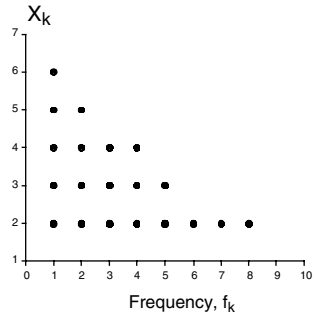
SUMMARY OF VARIOUS DATA PRESENTATIONS



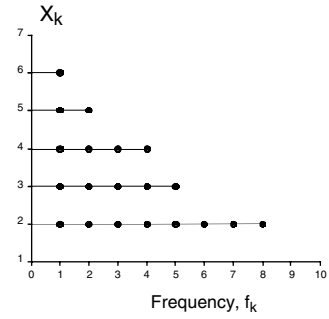
1. Chronicle plot of data



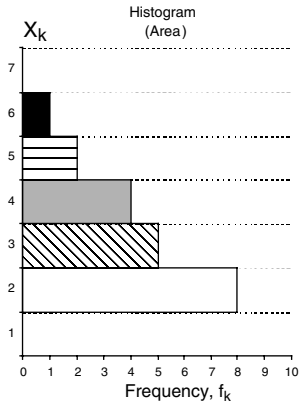
2. Data within cell intervals



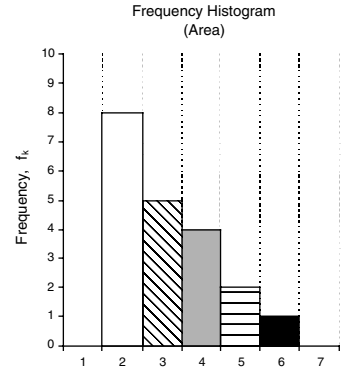
3. Frequency of data in cells



4. Bar graph of cell data



5. Histogram of cell data



6. Frequency histogram

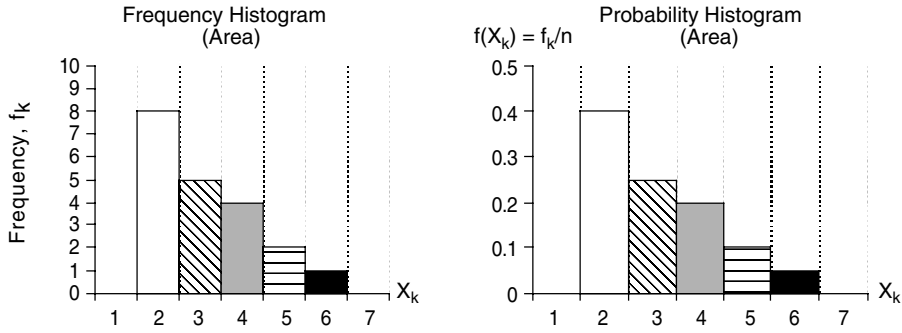
PROBABILITY DENSITY FUNCTION (PDF)

1. Problem: The frequency or number of observations within a cell, f_k, depends upon the total sample size, n.
   In the example, n = 20; however, if n is increased to 100, then values of frequency, f_k, may change.
2. Solution: Normalize frequency by the total sample size n.
   Defines the probability density function (pdf) of the grouped range values X_k:

$$f(X_k) = \frac{f_k}{n}$$

3. Assumes: Each observed sample is independent of others and represents equally likely events.
   Example Problem: n = 20

Cell, RV X_k	Frequency f_k	Probability $f(X_k) = f_k/n$
2	8	0.40
3	5	0.25
4	4	0.20
5	2	0.10
6	1	0.05
	20	1.00



MEAN OF FREQUENCY GROUPED DATA

Start with definition of mean. (Sum index *i* is over all n-observed *i*ndividual data values. Sum index *k* is over the number of *c*ell intervals, n_c.)

Sample Mean:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X(i) = \frac{1}{20}\sum_{i=1}^{20} X(i) = \frac{1}{20}\ [63]$$

$$= \frac{1}{20}[(2\cdot 8)+(3\cdot 5)+(4\cdot 4)+(5\cdot 2)+(6\cdot 1)]$$

$$= \frac{1}{20}\sum_{k=1}^{5} X_k f_k$$

$$= \frac{1}{n}\sum_{k=1}^{n_c} X_k f_k$$

OBSERVATIONS

1. Sum index "k" is over the number of cells, $n_c = 5$.
2. Divide the cell sum by the total of observations, n = 20.

MEAN OF PROBABILITY DENSITY FUNCTION

Assumes: All observed data are independent and equally likely.

$$\overline{X} = \frac{1}{20}\sum_{i=1}^{20} X(i) = \frac{1}{20}\sum_{k=1}^{5} X_k f_k$$

$$= \frac{1}{n}\sum_{i=1}^{n} X(i) = \frac{1}{n}\sum_{k=1}^{n_c} X_k f_k$$

$$= \sum_{k=1}^{n_c} X_k \frac{f_k}{n}$$

$$= \sum_{k=1}^{n_c} X_k f_k(X_k)$$

where the discrete pdf for each cell is defined as

$$f(X_k) = \frac{f_k}{n}$$

Check: (Determine sample mean using pdf)

$$\overline{X} = \sum_{k=1}^{n_c} X_k f(X_k) = \sum_{k=1}^{5} X_k f(X_k)$$

$$= 2 \cdot (0.40) + 3 \cdot (0.25) + 4 \cdot (0.20) + 5 \cdot (0.10) + 6 \cdot (0.05)$$

$$= 0.80 + 0.75 + 0.80 + 0.50 + 0.30$$

$$= 3.15 \text{ people}$$

FORMULAS FOR MEAN OR AVERAGE

1. Mean of all observed data (running average)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n_c} X(i)$$

where X(i) is the value of the i-th individual observation
n is the total number of observations
i is the sum index over all n-observed data values.

2. Mean of frequency grouped data (ranked and stacked)

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n_c} X_k f_k$$

where X_k is the nominal value of the k-th cell
f_k is the frequency of observations occurring in the k-th cell
n_c is the number of cells or data groups ($n_c \leq n$)

$$\sum_{k=1}^{n_c} fk = n$$

3. Mean of discrete probability density function (pdf)

$$\overline{X} = E[X_k] = \sum_{k=1}^{n_c} X_k f(X_k)$$

where $f(X_k) = \frac{f_k}{n}$ is probability of cell random variable X_k

$$\sum_{k=1}^{n_c} f(X_k) = 1.00$$

CUMULATIVE FREQUENCY FUNCTION

1. The number of observed data values that are below or equal to a stated value of X_k. (Think of a sieve.)
2. The stated value of X_k is often called the *threshold*: $T(X_k)$.
3. The number of observed data values *less than or equal to* this threshold are summed or "cumulated."

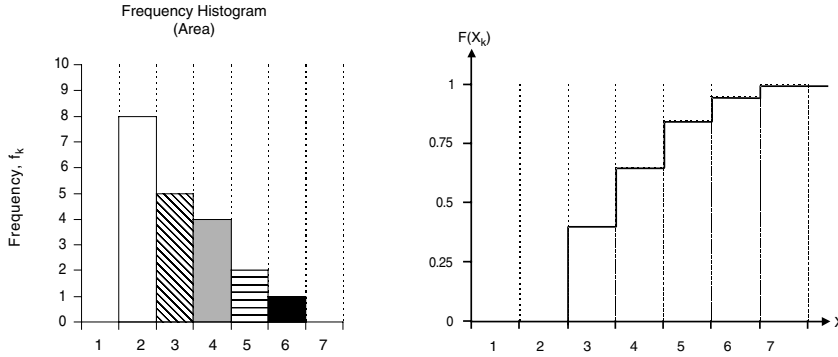
$$F_{nT} = \sum_{k=1}^{n_T} f_k$$

where n_T is the cell number $k(\leq n_c)$ set by the threshold $T(X_{nT})$.

Note: Cumulative distribution has the advantage that data do *not* need to be rank ordered before proceeding.

Example of Car Pool Observations

Cell (X_k)	Frequency	Threshold	Cumulation, F_k
$X_1 = 2$	$f_1 = 8$	$T(X_1)$	8
$X_2 = 3$	$f_2 = 5$	$T(X_2)$	8 + 5 = 13
$X_3 = 4$	$f_3 = 4$	$T(X_3)$	8 + 5 + 4 = 17
$X_4 = 5$	$f_4 = 2$	$T(X_4)$	8 + 5 + 4 + 2 = 19
$X_5 = 6$	$f_5 = 1$	$T(X_5)$	8 + 5 + 4 + 2 + 1 = 20



CUMULATIVE DISTRIBUTION FUNCTION (CDF)

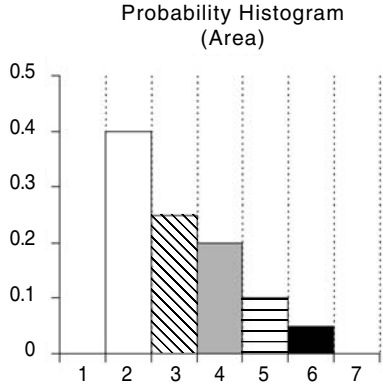
1. The number of observed data values below or equal to a stated value of X_k.
2. The stated value of X_k is often called the *threshold:* $T(X_k)$.
3. The probability of data values *less than or equal to* this threshold is the sum or "cumulation" of probabilities:

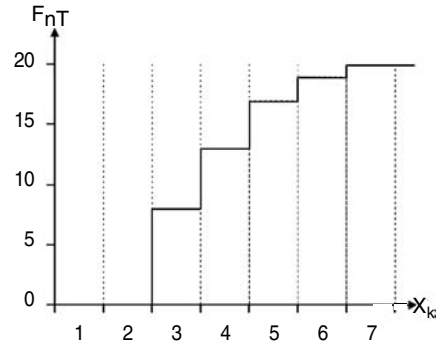
$$F(X_{nT}) = \sum_{k=1}^{n_T} f(X_k)$$

where n_T is the cell number n_c set by the threshold $T(X_{nT})$.

Example of Car Pool Observations

Cell (X_k)	Frequency	Threshold	Cumulation, $F(X_k)$
$X_1 = 2$	$f_1 = 0.40$	$T(X_1)$	0.40
$X_2 = 3$	$f_2 = 0.25$	$T(X_2)$	0.40 + 0.25 = 0.65
$X_3 = 4$	$f_3 = 0.20$	$T(X_3)$	0.40 + 0.25 + 0.20 = 0.85
$X_4 = 5$	$f_4 = 0.10$	$T(X_4)$	0.40 + 0.25 + 0.20 + 0.10 = 0.95
$X_5 = 6$	$f_5 = 0.05$	$T(X_5)$	0.40 + 0.25 + 0.20 + 0.10 + 0.05 = 1.0





PROBABILITY OF EXCEEDING THRESHOLD

Recall

$$\sum_{k=1}^{n_c} f(X_k) = 1.00$$

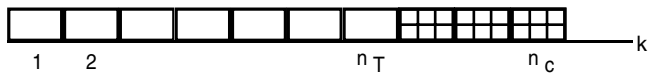
Then break total sum into two parts:

1. "less than or equal" plus "exceeds" equal "total"

$$\sum_{k=1}^{n_T} f(X_k) + \sum_{k=n_T+1}^{n_c} f(X_k) \equiv 1$$

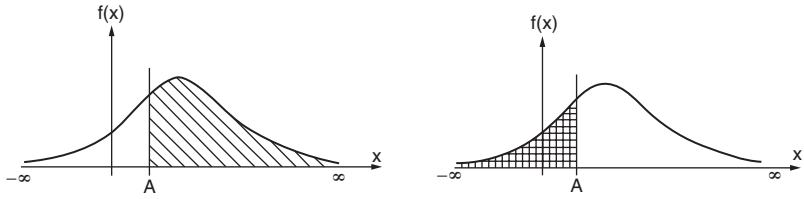
2. Probability of exceeding threshold:

$$\sum_{k=n_T+1}^{n_c} f(X_k) \equiv 1 - \sum_{k=1}^{n_T} f(X_k)$$



Continuous Probability

$$\int_{A}^{\infty} f(X)dX = 1 - \int_{-\infty}^{A} f(X)dX$$



DEVIATIONS OF DATA ABOUT MEAN

1. Deviation of the sample X(i) about the mean

$$d_i = (X(i) - \overline{X}) \text{ can be positive or negative}$$

2. Square of deviation of the sample X(i) about the mean

$$d_i^2 = (X(i) - \overline{X})^2 \text{ is always positive}$$

MEASURES OF DISPERSION

Sample Variance (Unbiased $\Rightarrow E(s^2) = \sigma^2$)

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X(i) - \overline{x})^2$$

$$= \frac{1}{(n-1)} \sum_{i=1}^{n} X(i)^2 - \frac{1}{(n-1)} \overline{x}^2$$

Note: the latter form is preferred for computations since computations can be done "on-the-run" as data is acquired.

Standard Deviation (Unbiased)

$$S = \left(\frac{1}{(n-1)} \sum_{i=1}^{n} (X(i) - \bar{x})^2 \right)^{1/2}$$

$$= \left(\frac{1}{(n-1)} \sum_{i=1}^{n} X(i)^2 - \frac{1}{(n-1)} \bar{x}^2 \right)^{1/2}$$

Recall

$$(a^2 \pm b^2)^{1/2} \neq a \pm b$$

PROBABILITY DENSITY AND EXPECTED VALUES

1. Probability density function $f(X_k)$

   Where $f(X_k) = \frac{f_k}{n}$ is the probability of the r.v.X_k
2. Expected value of sample mean (first moment about origin):

$$\overline{X} = E[X_k] = \sum_{k=1}^{n_c} X_k f(X_k)$$

Note: Summation of number of cells, k.

3. Sample variance (second moment about mean):

$$S^2 = E[(X_k - \overline{X})^2] = \sum_{k=1}^{n_c} (X_k - \overline{X})^2 f(X_k)$$

$$= E[(X_k)^2] - (E[X])^2 = \sum_{k=1}^{n_c} (X_k)^2 f(X_k) - \overline{X}^2$$

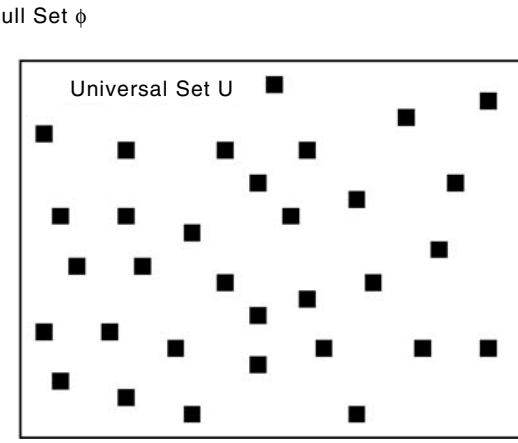
Special note: Statistics is concerned mainly with the mean and variance.

14 Set Theory

This chapter introduces the concepts of set theory, some general probability concepts, and some examples to facilitate the understanding of these concepts.

DEFINITIONS

1. Universal Set U
   Collection or aggregate of all possible elements
   Elements are outcomes or samples
2. Null Set ϕ
   Empty set with no elements

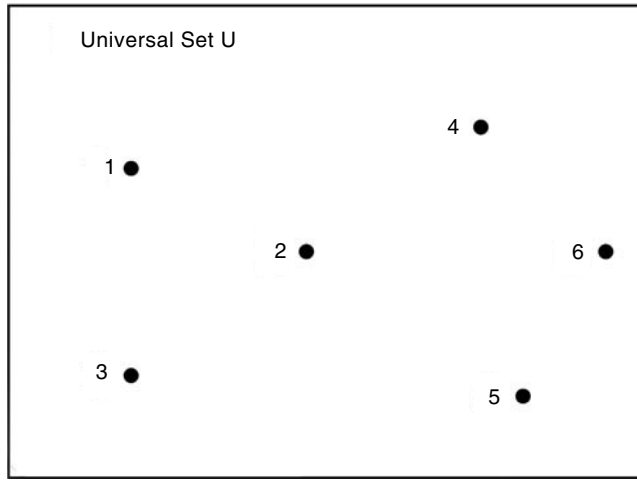


Probability $P(U) \equiv 1$
Probability $P(\phi) \equiv 0$

Example of Universal Set

Toss a single die, set of possible dots facing upward:

$$U = \{1, 2, 3, 4, 5, 6\}$$



SUBSETS OF ELEMENTS OF UNIVERSAL SET

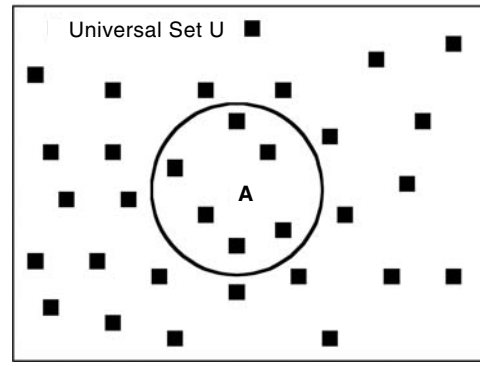
Subset A "contained in" or "element of" Universal Set U

$$A \subset U$$

Null Set ϕ

Probability:

$$0 \leq P(A) \leq 1$$



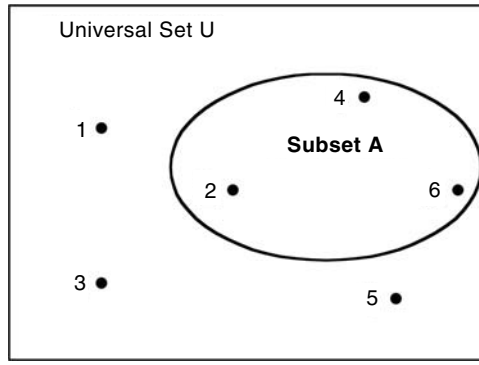
Example of Subset A of a Universal Set U

Toss a single die, the set of possible dots facing upward:

$$U = \{1, 2, 3, 4, 5, 6\}$$

Subset A is all *even valued* outcomes.

$$A = \{2, 4, 6\} \subset U = \{1, 2, 3, 4, 5, 6\}$$



"OR" SET OF OPERATION: UNION OF TWO SUBSETS

$A \cup B = C$

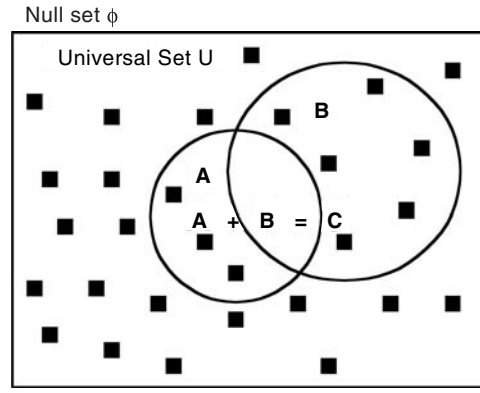
Set of all elements belonging together:

A *or* B *or* both

Sometimes written as "Sum"

$$A + B = C$$

(Related to failure probability of *parallel* systems)



"AND" SET OF OPERATION: INTERSECTION OF TWO SUBSETS

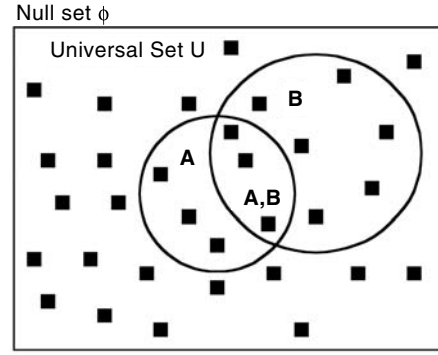
$A \cap B$

Set of those elements which belong to both A *and* B

Sometimes written as "Product"

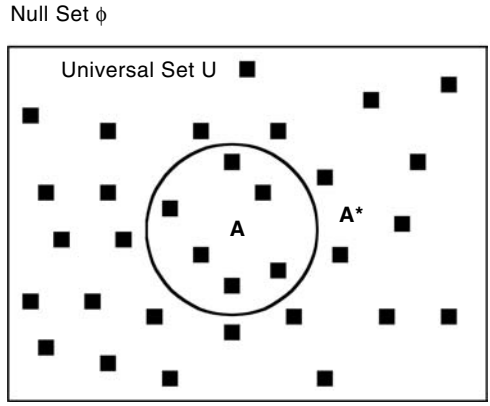
A, B or $A \cdot B$ or AB

(Related to failure probability of *series* systems)



COMPLEMENTARY SET, A* (OTHER NOTATION: $\overline{A}$, A')

The set of all other outcomes excluding those in subset A



Union	$A \cup A^* = U$	Universal set
Intersection	$A \cap A^* = 0$	Null set
Probability	$P(A) = 1 - P(A^*)$	

DE MORGAN'S LAWS OF COMPLEMENTS

$$(A \cap B)^* = A^* \cup B^*$$

$$(A \cup B)^* = A^* \cap B^*$$

Then:

1. $A = (A \cap B) \cup (A \cap B^*)$
2. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

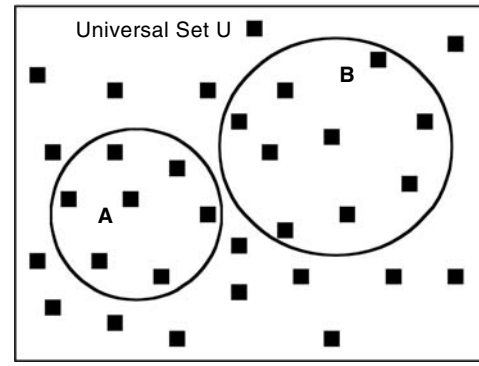
"DISJOINT" SETS (MUTUALLY EXCLUSIVE EVENTS)

1. No intersection of set A and set B

   $A \cap B = \phi$

   No common elements in A and B

Null Set ϕ

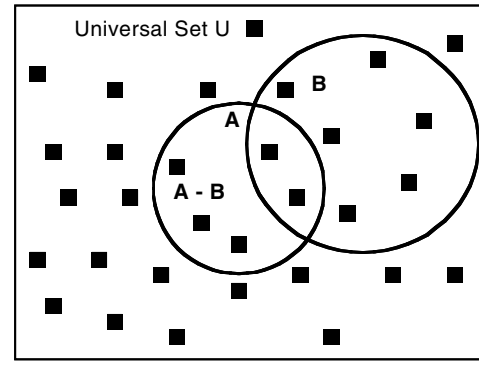


2. "Difference" set
   Intersection of set A and the complement of set B

   $A - B = A \cap B^*$

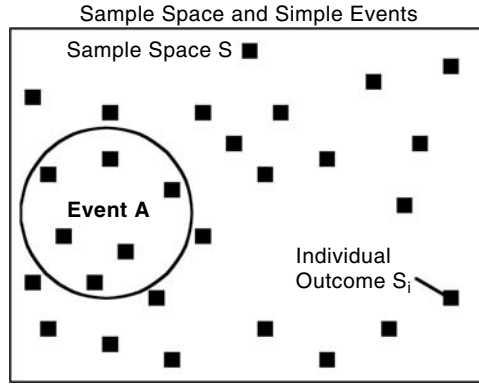
   Subsets of those elements of A that do not belong to B.

Null Set ϕ

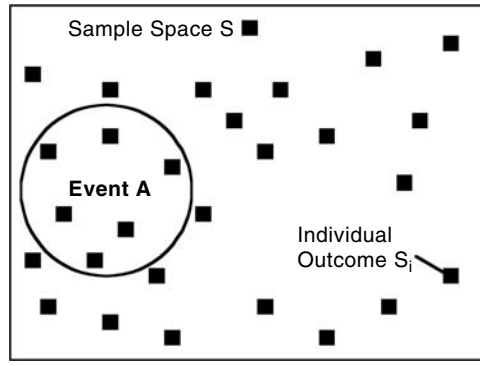


SAMPLE SPACE: S

1. Equivalent to universal space U
   - Related to a random experiment
     - Uncertainty of results or outcomes
   - Consists of all possible outcomes or samples
   - Each output is a simple or elementary event, S_i
   - Discrete, e.g., selecting a card or coin tosses
   - Continuous, e.g., diameter or weight of rods
   - Finite or infinite



2. Events of sample space: A
   - Simple or elementary event is one (individual) outcome of sample space, S_i.



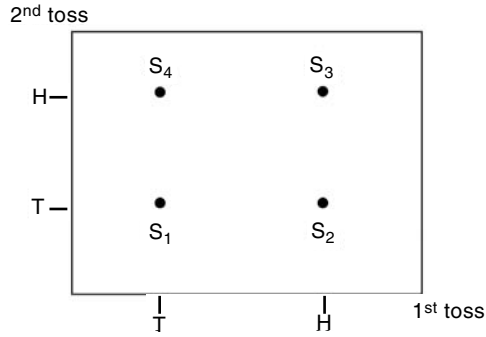
   - An event A is a subset or grouping of simple events; e.g., event A of n simple events: $A = \{S_1, \ldots, S_n\}$
   - Elements of event A are samples with defined characteristics; e.g., the odd numbers tossed on dice.

If the outcome of an experiment S_i is an element of A, then "the event A has occurred."

Examples of Sets

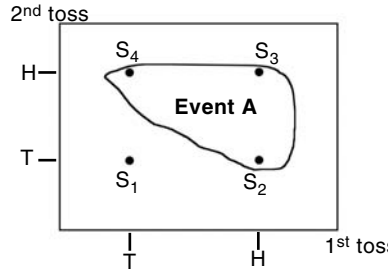
Given: Random experiment in a twice-tossed coin.

1. Find the sample space S of the four possible outcomes:
   Simple events: $S_1 = TT$, $S_2 = HT$, $S_3 = HH$, $S_4 = TH$
   Sample space: $S = (S_1, S_2, S_3, S_4)$



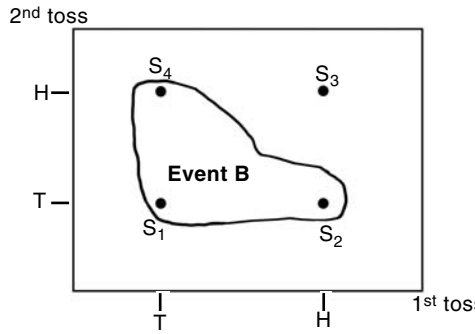
2. Find the subset or event A defined as when: At least one head occurs.

$$\text{Event A} = (\text{HT, HH, TH}) = (S_2, S_3, S_4)$$



3. Find the subset or event B defined as when: At least one tail occurs.

$$\text{Event B} = (\text{TT, HT, TH}) = (S_1, S_2, S_4)$$

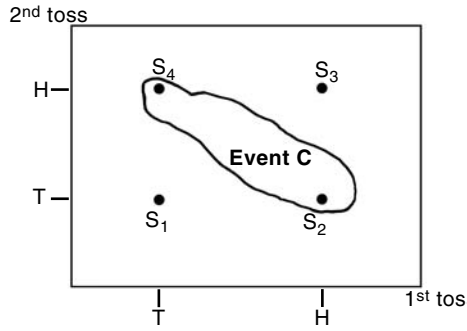


4. Find the subset or event C defined as when: At least one head (event A) AND one tail (event B) occur.

$$\text{Event C} = (\text{HT, TH}) = (S_2, S_4)$$

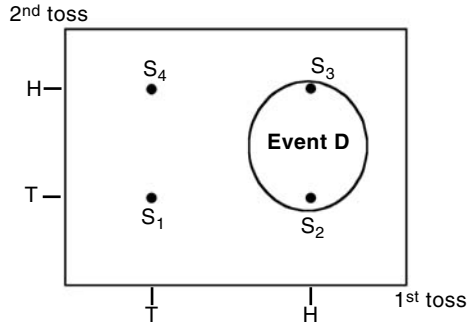
"AND" in set theory is an "intersection," meaning "both A and B."

$$C \equiv A \cap B$$
$$\equiv (S_2, S_4)$$



5. Find the subset or event D defined as when: The first toss is a head.

$$\text{Event D} = (\text{HT, HH}) = (S_2, S_3)$$



6. Find the subset or event E defined as when: At least one head (event A) OR at least one tail (event B) occurs.

$$\text{Event E} = (\text{TT, HT, HH, TH}) = (S_1, S_2, S_3, S_4)$$

"OR" in set theory is "union," meaning "either A *or* B *or* both."

$$E \equiv A \cup B = (S_1, S_2, S_3, S_4) = S$$

7. Find the event F that CANNOT occur assuming the event A does occur. Event A is defined as when at least one head occurs.

$$\text{Event A} = (\text{HT, HH, TH}) = (S_2, S_3, S_4)$$

Event F is the event "not A."

$$\text{Event F} = (\text{TT}) = (S_1)$$

Event A and event F are "mutually exclusive" or "disjoint" in set theory.

$$A \cap F = \phi = A \cap A'$$

Hence the event

$$F = A' = (S_1) = (\text{TT})$$

PROBABILITY CONCEPTS

- Probability quantifies the chances an event will occur.
- Bound probability of an event A.

$$0 \leq P(A) \leq 1$$

- Probability of an event A that is certain to occur.

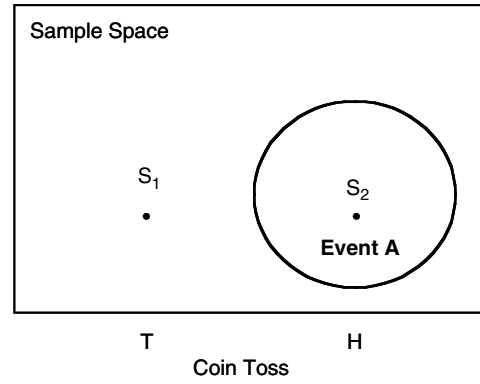
$$P(A) = 1 \qquad [100\%]$$

- Probability of an event A that cannot occur.

$$P(A) = 0 \qquad [0\%]$$

1. Estimates of probability
   a. **Classical or *a-priori* approach** —This approach is based on "equally likely" events.
      - Assumes all simple events or individual outcomes are "equally likely."
      - Sample space S: A random experiment with a total of n possible outcomes.
      - Defined event A: Has r of these possible outcomes [$r \leq n$].
      - Probability of event A: $P(A) = \frac{r}{n}$

Example 1



Random experiment: Single toss of one coin.

Find: Probability of a head.

Sample space: $S = \{S_1, S_2\} = \{T, H\}$.

TWO mutually exclusive events: n = 2.

Event A: Tossing ONE head S_2; r = 1.

Assume each outcome is "equally likely."

Probability of Event A = S_2: $P(A) = P(S_2) = r/n = 1/2$

Example 2

Random experiment: Twice-tossed coin.

Find: Probability of two heads.

Sample space: $S = \{S_1, S_2, S_3, S_4\} = \{TT, TH, HT, HH\}$

FOUR mutually exclusive events: n = 4.

$$S_1 = TT, S_2 = TH, S_3 = HT, S_4 = HH$$

Event A: Tossing TWO heads, which is simple event S_4; r = 1.

Assume the four possible outcomes are "equally likely."

Probability of Event A = S_4: $P(A) = P(S_4) = r/n = 1/4$

Example 3

Random Experiment: Single roll of fair die.

Find: Probability of rolling a 3.

Sample space: $S = \{S_1, S_2, S_3, S_4, S_5, S_6\} = \{1, 2, 3, 4, 5, 6\}$

SIX mutually exclusive simple events: n = 6.

Event A: Rolling the ONE number 3 on the die (i.e., S_3); r = 1.

Assume each of the six possible outcomes is "equally likely."

Probability of Event A = S_3: $P(A) = P(S_3) = r/n = 1/6$

b. **Frequency or *a-posteriori* approach —** This approach is based on a "very large" number of independent samples.
   - Assumes after n repetitions of a given random experiments, a simple event S_i is observed to occur in r_i of these.
   - Assumes n is "very large."
   - Vague about how "large" n must be.
   - Empirical probability: $P(A) = P(S_i) \approx \frac{r_i}{n}$

Example 4

Random Experiment: Tossing a single coin.

Sample space: $S = \{S_1, S_2\} = \{T, H\}$

TWO mutually exclusive events: n = 2

Event A: The simple event S_2 (head) will occur.

Observations: After n = 1000 tosses we find we have accumulated 483 heads and 517 tails.

Empirical probability of the event "a head:"

$$P(A) = P(S_2) \approx \frac{483}{1000} = 0.483$$

Special note: This is not the same as the 0.5 of the classical approach (Example 1).

c. **Axiomatic approach —** This approach is based on set theory.
Rule 1: Probability of an event A is bounded.

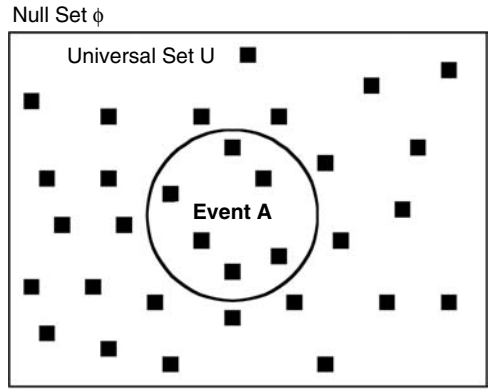
$$0 \leq P(A) \leq 1$$

Certainty or the sure occurrence of event A
Using Sets: Event A = Sample Space S

$$P(A) = P(S) = 1$$

Impossibility or absence of the event A
Using Sets: Event A = Null Space ϕ

$$P(A) = P(\phi) = 0$$



Rule 2: Total or cumulative probability

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

"OR"

$P(A + B) = P(A) + P(B) - P(A \cdot B)$

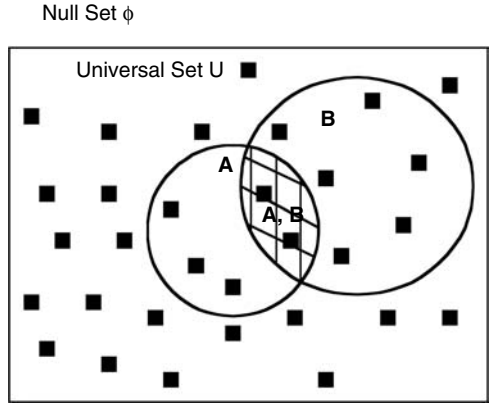
$P(A + B)$ = probability of either events A *or* B *or* both

$P(A)$ = probability of event A alone

$P(B)$ = probability of event B alone

$P(A \cap B)$ = probability of both events A *and* B

"AND" implies intersection of sets: $(A \cap B) \equiv (A \cdot B)$, in other words, it removes one count of common simple events.



Useful alternate form: $P(A + B) = 1 - P(A^* B^*)$

Similarly, with THREE EVENTS (parallel components of "OR" events)

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

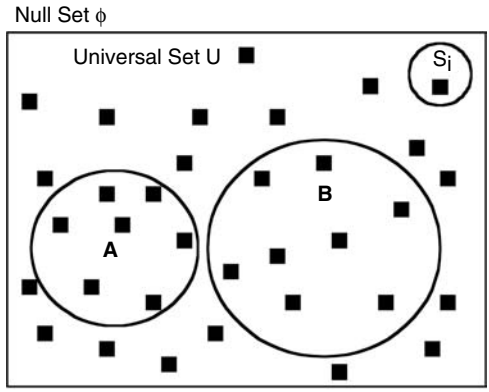
Rule 2A: Total probability of MUTUALLY EXCLUSIVE EVENTS (not to be confused with independent events). In this case, the occurrence of one event precludes the occurrence of another.

Sets: Events A and B are mutually exclusive if they have no common simple event.

"AND" intersection set is zero: $A \cap B = 0$

Hence, P(A + B) = Probability of either events A *or* B *or* both becomes the sum of the individual probabilities:

$$\begin{aligned} P(A + B) &= P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) \end{aligned}$$



Note: All individual simple events, S_i, are mutually exclusive.

EXAMPLE 5

Random Experiment: A single toss of a fair die.

Sample space: Six "equally likely" simple events

$S_1 = 1$, $S_2 = 2$, $S_3 = 3$, $S_4 = 4$, $S_5 = 5$, $S_6 = 6$

Event A: Defined as either a 2 *or* 5 will occur.

$A = \{S_2, S_5\}$

Event B: Defined as any even number will occur.

$B = \{S_2, S_4, S_6\}$

Probability of individual event:

$P(S_i) = 1/6$; I = 1, 2, … 6

Probability of Event A:

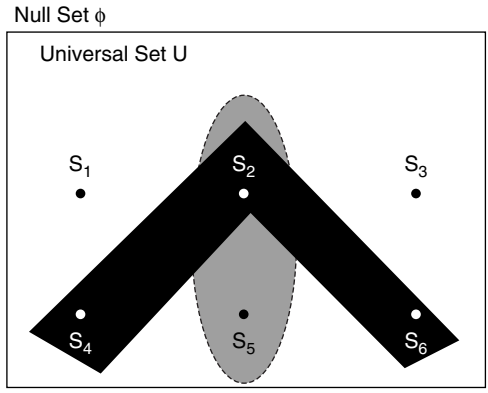
$P(A) = P(S_2 + S_5) = P(S_2) + P(S_5)$

$$\frac{1}{6}+\frac{1}{6}=\frac{1}{3}$$

Probability of Event B:

$P(B) = P(S_2 + S_4 + S_6) = P(S_2) + P(S_4) + P(S_6)$

$$\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{1}{2}$$



EXAMPLE 5A

Random Experiment: A single toss of a fair die.

Find: Total probability that the events A *or* B *or* both occur.

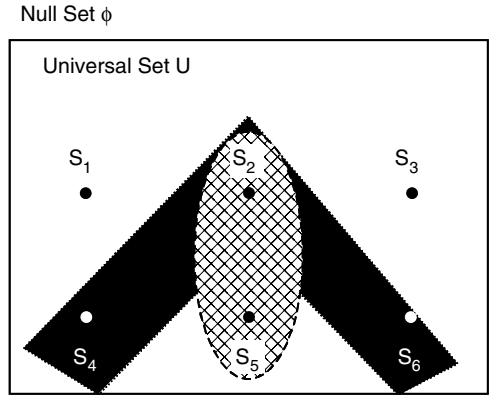
"OR" total probability

$$P(A + B) = P(A) + P(B) – P(A, B)$$

where $P(A) = 1/3$; $P(B) = 1/2$

P(A, B) = probability of both events A *and* B = $P(S_2) = 1/6$

$P(A + B) = 1/3 + 1/2 – 1/6 = 4/6 = 2/3$



Note: Total probability must be greater or equal than that of A or B alone.

Rule 3: Joint and Conditional Probability

Joint probability: Probability that an outcome will satisfy both events A and B simultaneously.

$$P(A \bullet B) = P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$$

Conditional Probability: Probability of A given B

P(A | B) = Probability of A *given that* B has occurred

$$P(A \mid B) = \frac{P(A \cdot B)}{P(B)} \equiv \frac{P(A \cap B)}{P(B)}$$

Rule 3A: Joint probability of mutually exclusive events. Mutually exclusive is not the same as independent.

Joint probability of two mutually exclusive events A and B:

$$P(A \cdot B) \equiv P(A \cap B) = 0$$

Conditional probability of two mutually exclusive events is undefined since by definition the occurrence of event B excludes the occurrence of event A (see Bayes' Rule).

For THREE EVENTS (multiplication)

$$P(A \cdot B \cdot C) = P(A \cap B \cap C) = P(A)P(B \mid A)P(C \mid A \cap B) = P(A)P(B)P(C) \text{ if independent}$$

Rule 4: Independent events. Events such that the occurrence of one has no effect on the occurrence of another.

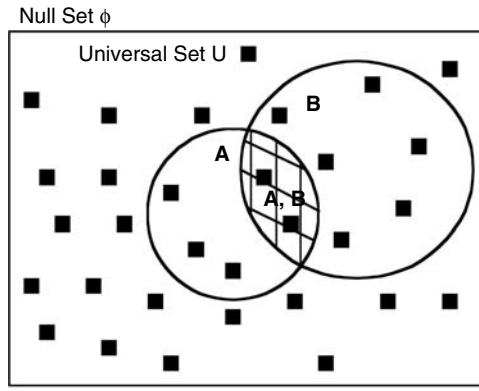
Conditional probability of independent events:

$$P(A \mid B) = P(A)$$

Joint probability of independent events:

$$\begin{aligned} P(A \cdot B) = P(A \cap B) &= P(A)P(B \mid A) \\ &= P(A)P(B) \end{aligned}$$

This actually turns out to be the definition of "independence" of events.



The reader should note that since both P(A) and P(B) are less than unity, their product will be smaller than either.

$$(\text{e.g., } 1/4 \times 1/3 = 1/12)$$

Therefore, the total probability of independent events may be shown as:

$$\begin{aligned} P(A \cup B) \equiv P(A) + P(B) - P(A \cap B) \\ = P(A) + P(B) - P(A)P(B \mid A) \\ = P(A) + P(B) - P(A)P(B) \end{aligned}$$

Example 5B

Random Experiment: A single toss of a fair die

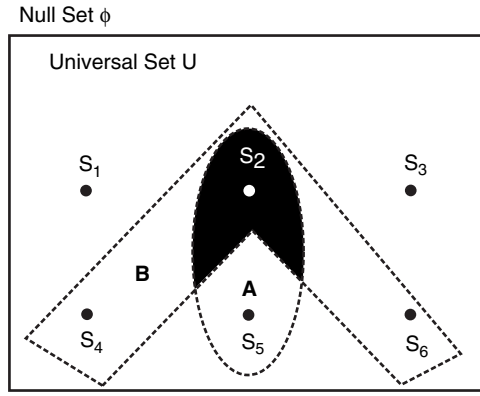
Find: Joint probability that events A *and* B occur

$$P(A \cdot B) = P(A)P(B \mid A)$$

Independent Events: Events A and B are independent since each consists of a set of the independent sample space S.

Mutually Exclusive Events: Events A and B are not mutually exclusive since they have one common simple event S_2.

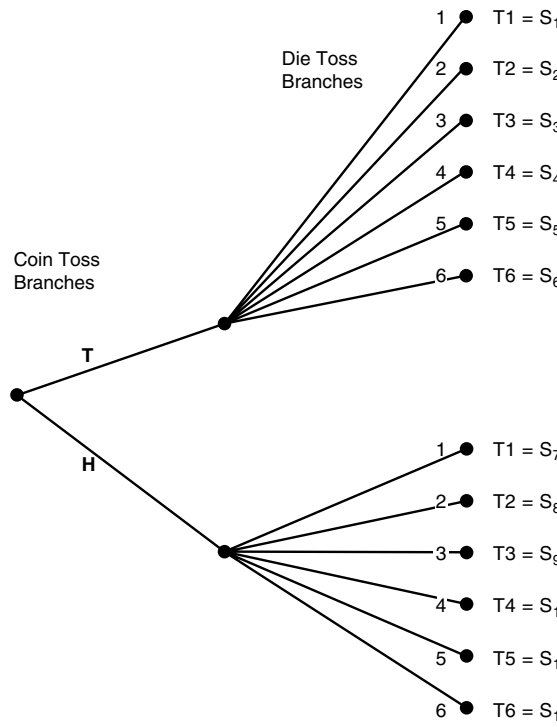
$$P(A \cdot B) = P(A)P(B \mid A) = P(A)P(B)$$
$$= 1/3 \times 1/2 = 1/6$$



EXAMPLE 6

Random Experiment: A single toss of a fair die and a coin.

Sample Space: Twelve "equally likely" simple events.



Sample Space S: Twelve "equally likely" simple events

$$P(S_i) = 1/12; \quad i = 1, 2, \ldots, 12$$

Event A: Coin is a head and die has an even number

$$A = \{S_8, S_{10}, S_{12}\}$$

$$P(A) = P(S_8) + P(S_{10}) + P(S_{12}) = 1/12 + 1/12 + 1/12 = 1/4$$

Event B: Any coin toss and die less than 5 (i.e., 1, 2, 3, 4)

$$B = \{S_1, S_2, S_3, S_4, S_7, S_8, S_9, S_{10}\}$$

$$P(B) = 8P(S_i) = 8/12 = 2/3$$

Example 6A

Random Experiment: A single toss of a fair die and a coin.

Find: Joint probability of independent events A and B.

Joint probability that A *and* B occur

$$P(A \cdot B) = P(A)P(B \mid A)$$

Independent Events: Events A and B are assumed independent; the probability of event B given that A has occurred is simply

$$P(B \mid A) = P(B)$$

Joint probability of the given independent events is

$$P(A \cdot B) = P(A)P(B) = 1/4 \cdot 2/3 = 2/12 = 1/6$$

Note: Since both events must occur, probability of "success" will be *smaller* than the probability of either event separately.

Independent events imply

$$P(B|A) \equiv \frac{P(A \cdot B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B)$$

Hence, the occurrence of event A has no influence on the probability of event B occurring.

Rule 5: Bayes' Rule

If the events A_1, A_2, …, A_n are mutually exclusive whose sum (union) form the complete sample space S

$$A_1 + A_2 + \ldots + A_n = S$$

then one of the events A_i must occur.

If any other event B of the space S occurs, then the probability of the event A_m to cause the event B is given by Bayes' Rule:

$$P(A_m \mid B) = \frac{P(A_m)P(B \mid A_m)}{P(A_1)P(B \mid A_1) + \ldots + P(A_n)P(B \mid A_n)}$$

Bayes' Rule is referred to as the "probability of causes," or a "conditional probability."

2. Complementary events (See also Binomial Distribution)

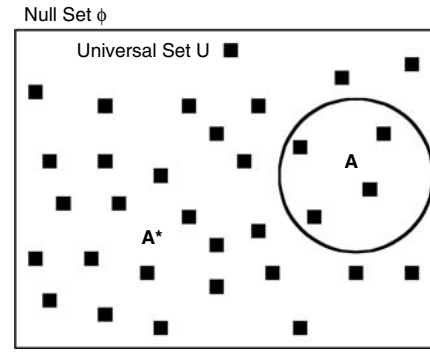
If event A represents a "success," then the complement of event A, denoted A*, represents "failure."

Probability of event A or a "success:"

$$P(A) = p$$

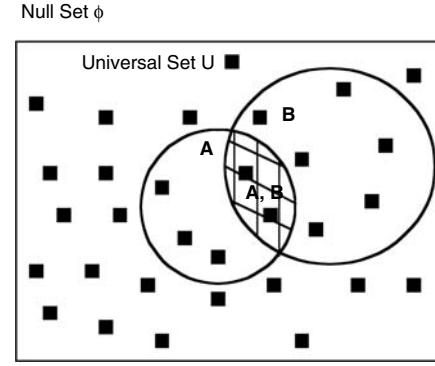
Probability of not event A or a "failure:"

$$P(A^*) = 1 - P(A) = q$$



Total probability

Probability of event A or event B or both a "success:"



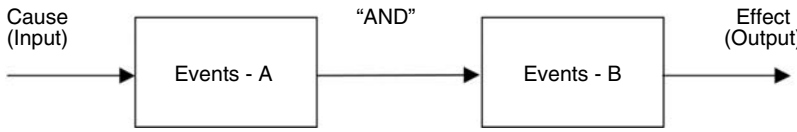
$$P(A+B) \equiv P(A)+P(B)-P(A\cdot B)$$
$$= 1-P(A^*\cdot B^*)$$
$$= 1-P(A^*)\cdot P(B^*), \text{ if independent}$$

Probability of not A "AND" not B: (where not A = A*)

$$P(A^*\cdot B^*) = 1 - P(A + B)$$

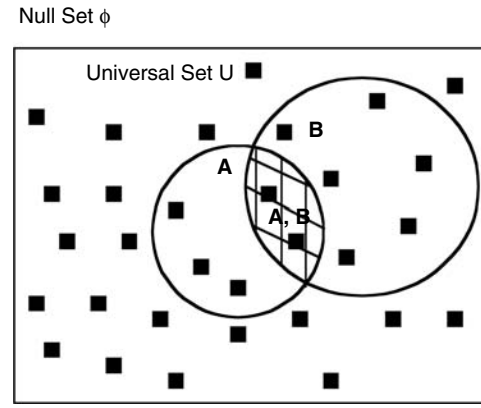
3. Series system of events
   a. No redundancy (A chain is as strong as its weakest link.)

Product rule for individual independent probabilities (Series components):



Joint Probability:

$P(A \cap B) = P(A \cdot B)$



$$P(A\cdot B) = P(A)P(B \mid A)$$
$$= P(A)P(B), \text{ if independent}$$

Note:

1. Independent is not the same as "mutually exclusive."
2. Individual probabilities are always less than unit.
3. Product probability for the series is always *smaller* than lowest individual series components.

Numerical Example

Components A and B represent events both of which must be satisfied simultaneously for system function. Past history indicates that the individual probability of these components to function properly is:

$$P(A) = 0.90 \text{ and } P(B) = 0.80$$

The probability that the system will function properly is then the probability of component A "AND" the probability of component B.

Since A and B are assumed to be independent events, the probability of the series is

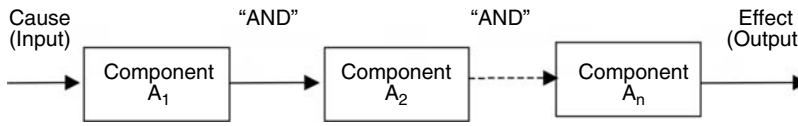
$$\begin{aligned} P(A \cdot B) &= P(A)P(B \mid A) \\ &= P(A)P(B) \\ &= 0.9 \cdot 0.8 \\ &= 0.72 \end{aligned}$$

In any series system,

The probability of the series system is less than those of either of the individual components.

In the series configuration, if any of the components fail the entire system fails.

b. Series components



A system consists of a series of n components (no redundancy).
Successful operation of each component is independent (independent implies components do not interact).
Reliability of individual components:

$$R_i = P(A_i) = p_i$$

Unreliability of individual components:

$$Q_i = P(A^*_i) = 1 - p_i = q_i$$

System reliability is the joint probability of all components:

$$\begin{aligned} R_s &\equiv P(A_1 \cdot A_2 \ldots A_n) \\ &= P(A_1)P(A_2 \mid A_1)\ldots P(A_n \mid A_1A_2 \ldots A_{n-1}) \\ &= P(A_1)P(A_2)\ldots P(A_n), \text{ if independent} \\ &= \prod_{i=1}^{n} P(A_i) \\ &= \prod_{i=1}^{n} R_i \end{aligned}$$

Note: Serial system reliability is product of individual reliability of the components R_i. Series systems are inherently *less* reliable.

System unreliability:

$$Q_s = 1 - R_s = 1 - \prod_{i=1}^{n} R_i = 1 - \prod_{i=1}^{n} (1 - Q_i)$$

Example

A system consists of a series of three identical switches.

Assume 1: Each switch has the same probability not to fail of p.

Assume 2: Performance probability of each switch is independent.

System reliability: Probability of system not failing

$$\begin{aligned} R_s &\equiv P(A_1, A_2, A_3) \\ &= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2) \\ &= P(A_1)P(A_2)P(A_3), \text{ if no interaction} \\ &= p \cdot p \cdot p = p^3 \end{aligned}$$

System unreliability: $Q_s = 1 - R_s = 1 - p^3$

Numerical Example

Assume each switch has a reliability or probability of "successful" performance of 90%.

$$P(A_i) = R_i = p = 0.90$$

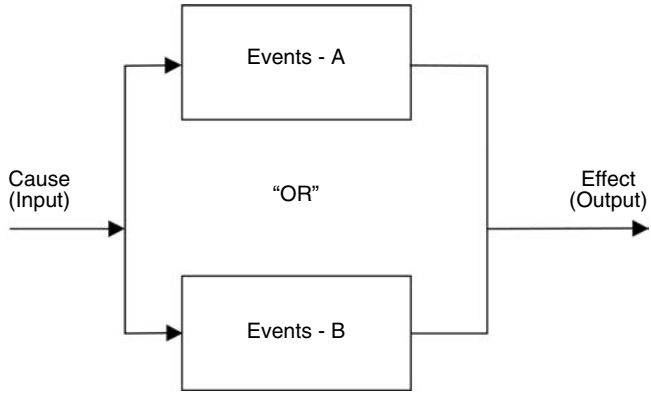
System Reliability: $R_s = 0.90 \cdot 0.90 \cdot 0.90 = 0.729$

Unreliability of individual switch: $Q_i = 1 - R_i = 0.10$

System unreliability:

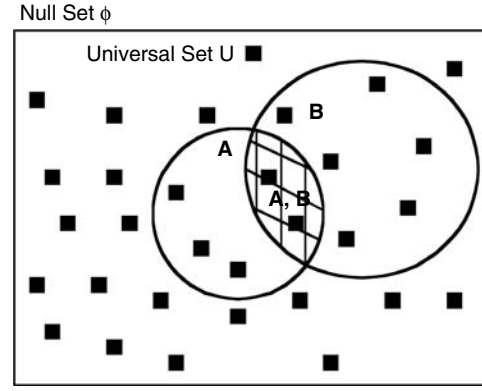
$$Q_s = 1 - R_s = 1 - p^3 = 1 - (0.90)^3 = 1 - 0.729 = 0.271$$

4. Parallel system of events — some form of redundancy
   Total or cumulative rule for individual independent probabilities:



Total Probability:

$$P(A \cup B) = P(A + B)$$



$$P(A + B) \equiv P(A) + P(B) - P(A \cdot B)$$
$$= P(A) + P(B) - P(A)P(B \mid A)$$
$$= P(A) + P(B) - P(A) \cdot P(B), \text{ if independent}$$

Note:

1. Individual probabilities are always less than unity.
2. Because of redundancy, total probability for the parallel system can be greater than the smallest component.

Numerical Example

Components A and B represent events both of which must be satisfied simultaneously for a system function. Past history indicates that the individual probability of these components to function properly is:

$$P(A) = 0.90 \quad \text{and} \quad P(B) = 0.80$$

The probability of the *system* to function properly is then the probability of component A "OR" the probability of component B.

Since A and B are assumed to be independent events, the probability of the parallel system is:

$$\begin{aligned} P(A + B) &\equiv P(A) + P(B) - P(A \cdot B) \\ &= P(A) + P(B) - P(A) \cdot P(B) \\ &= 0.90 + 0.80 - 0.90 \cdot 0.80 \\ &= 1.70 - 0.72 \\ &= 0.98 \end{aligned}$$

The reader should notice that the redundancy of the parallel system that allows either-or components to function results in a system that remains functional with higher probability than the probability of either of the individual components acting alone. Therefore, redundancy increases reliability, which means that a system of n components is connected in parallel. Another way of looking at this is:

a. Successful operation of each component is independent. Again, independent here implies components do not interact.
b. Each component has a reliability $P(A_i) = p_i$; failure probability of each component is $P(A^*_i) = Q_i = 1 - p_i = q_i$

Pictorially this may be shown in Figure 14.1.

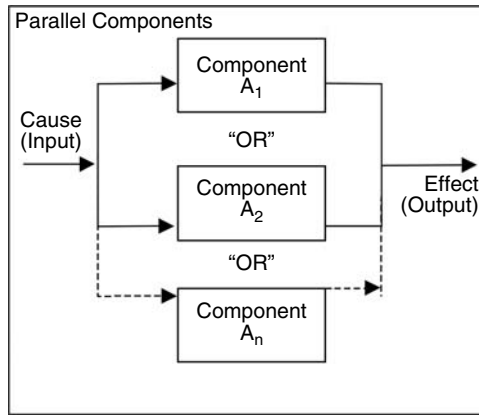


FIGURE 14.1 Parallel components.

We can see then from Figure 14.1 that the System Reliability is the total probability of all components. This also may be shown as:

$$R_p \equiv P(A_1 + A_2 + \ldots + A_n)$$

$$= 1 - P(A^*_1 \cdot A^*_2 \cdot \ldots \cdot P(A^*_n))$$

$$= 1 - P(A^*_1)P(A^*_2)\ldots P(A^*_n), \text{ if independent}$$

$$= 1 - \prod_{i=1}^{n} P(A^*_i)$$

$$= 1 - \prod_{i=1}^{n} Q_i$$

System unreliability:

$$Q_p = \prod_{i=1}^{n} Q_i$$

General Example A

In the previous example, we had: P(A) = 0.90 and P(B) = 0.80. If this is a parallel system then

$$R_p = 1 - P(A^*)P(B^*) = 1 - 0.10 \cdot 0.20 = 1 - 0.02 = 0.98$$

General Example B

A system consists of 3 identical switches in parallel.

Assume 1: Each switch has the same probability not to fail of p.

Assume 2: Performance probability of each switch is independent.

System Reliability: Probability of system not failing: (q = 1 – p)

$$R_p \equiv P(A_1 + A_2 + A_3)$$

$$= 1 - P(A^*_1\, A^*_2\, A^*_3)$$

$$= 1 - q \cdot q \cdot q = 1 - q^3$$

System Unreliability: $Q_p = 1 - R_p = 1 - (1 - q^3) = q^3$

Numerical Example

Assume each switch has a reliability or probability of "successful" performance of 90%, which corresponds to 10% "failure."

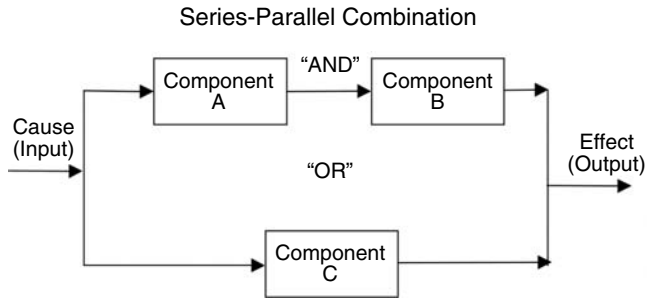
Individual Switch Reliability: $P(A_i) = R_i = p = 0.90$

Individual Switch Unreliability: $P(A^*_i) = Q_i = q = 0.10$

System Reliability: $R_p = 1 - 0.1 \cdot 0.1 \cdot 0.1 = 1 - 0.001 = 0.999$

System Unreliability: $Q_p = 1 - R_p = q^3 = (0.1)^3 = 0.001$

Example 1: Simple Combination of Series-Parallel System



Probability of success for individual components: P(A), etc.

Probability of failure for individual components: P(A*), etc.

Probability of success in series branch ("AND") if independent:

$$P(AB) = P(A)P(B) = 1 - P(A^* + B^*)$$

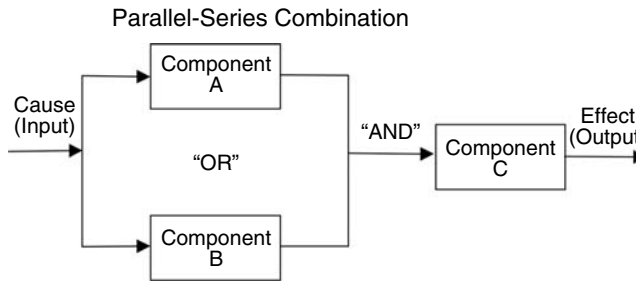
Reliability is success probability of series-parallel system shown:

$$\begin{aligned} R_{sys} \equiv P(AB + C) &= P(AB) + P(C) - P(AB)P(C) \\ &= P(AB)[1 - P(C)] + P(C) \\ &= \{1 - P(A^* + B^*)\}[P(C^*)] + P(C) \\ &= P(C^*) - P(A^* + B^*)P(C^*) + P(C) \\ &= P(C^*) + P(C) - P(A^* + B^*)P(C^*) \\ &= 1 - P(A^* + B^*)P(C^*) \end{aligned}$$

If probability of all components equal: $P(A) = P(B) = P(C) = p$

$$R_{sys} = 1 - (1 - p^2)(1 - p) = 1 - q^3$$

Example 2: Simple Combination of Parallel-Series System



Probability of success for individual components: P(A), etc.

Probability of failure for individual components: P(A*), etc.

Probability of success in parallel circuit ("OR") if independent:

$$P(A + B) = P(A) + P(B) - P(A)P(B)$$

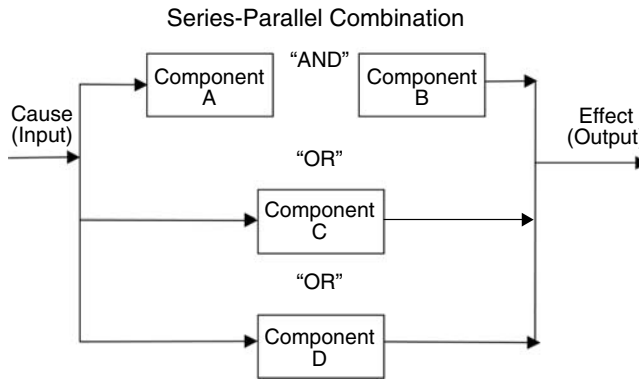
Reliability is success probability of parallel-series system shown:

$$\begin{aligned} R_{sys} \equiv P([A + B]C) &= P(A + B)P(C) \\ &= [1 - P(A^* B^*)]P(C) \\ &= [1 - P(A^*)P(B^*)]P(C) \\ &= 1 - [1 - P(A)(1 - P(B))]P(C) \end{aligned}$$

If probability of all components equal: P(A) = P(B) = P(C) = p

$$R_{sys} = [1 - 1 - p^2]p = [2p - p^2]p = [1 - q^2]p$$

Example 3: Simple Combination of Parallel-Series System



Reliability is success probability of series-parallel system shown:

$$R_{sys} = P(AB + (C + D)) = P(AB) + P(C + D) - P(AB)P(C + D)$$

Joint and total probabilities can be expressed in term of individual probabilities.

Series branch has joint probability: P(AB) = P(A)P(B)

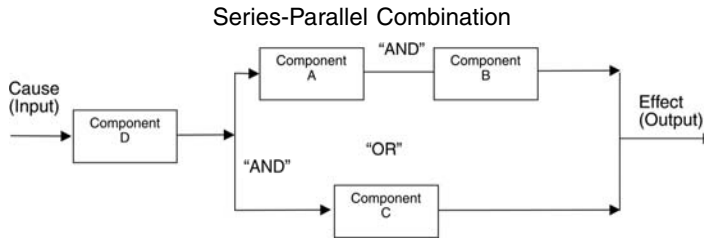
Parallel branches have total probability:

$$P(C + D) = P(C) + P(D) - P(C)P(D)$$

If probability of all components equal: P(A) = P(B) = P(C) = P(D) = p

$$R_{sys} = p^2 + p + p - p^2 - p^2(p + p - p^2) = 2p - 2q^3 + p^4$$

Example 4: Simple Combination of Parallel-Series System



Reliability is success probability of series-parallel system shown:

$$\begin{aligned} R_{sys} &\equiv P[D \cdot (AB + C)] \\ &= P(D)P[(AB) + C] \\ &= P(D)[P(AB) + P(C) - P(AB)P(C)] \\ &= P(D)[P(A)P(B) + P(C) - P(A)P(B)P(C)] \end{aligned}$$

Note: This form may be more convenient to use than that developed in Example 1 where we introduced the complement probabilities.

If all probability of all components equal: P(A) = P(B) = P(C) = p

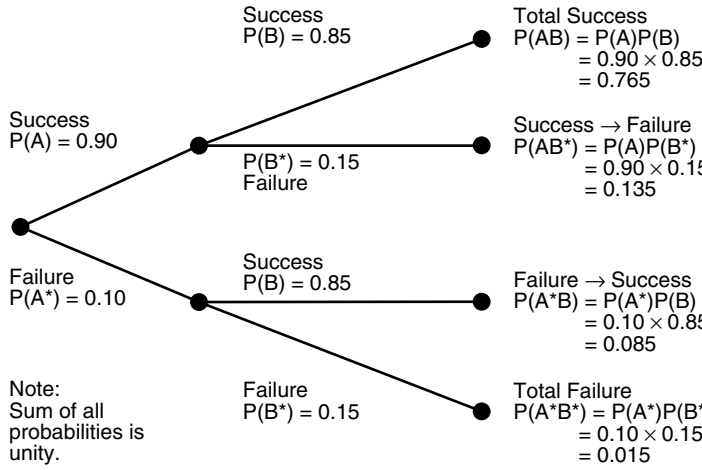
$$R_{sys} = p[p \cdot p + p - p \cdot p \cdot p] = p^3 + p^2 - p^4$$

5. Sequence tree diagram

A tree diagram is very useful in determining decisions where probabilities are known and they are of two options. An example adapted from O'Connor (1996) will illustrate the point. The reliability of missile A to hit target is 90%; that of missile B is 85%.

A salvo of both missiles is launched. Determine the possibility of at least one hit.

The tree diagram indicates that four mutually exclusive outcomes may occur:



Probability of at least one hit:

$$P(AB) + P(AB^*) + P(A^*B) = 0.765 + 0.135 + 0.085 = 0.985$$

REFERENCE

O'Connor, P., *Practical Reliability Engineering*, John Wiley & Sons, New York, 1996.

15 Permutations and Combinations

This chapter discusses elementary probability calculations and the "counting rules" that underlie the development of several distributions such as the binomial, hypergeometric, and Poisson. Perhaps one of the most essential concepts in statistical theory as applied in the field of quality is the issue of combinations. This chapter presents a cursory overview of this mathematical expansion and its use. If you are interested in pursuing the subject in more detail, please consult some basic statistics or probability books.

RULES

We begin by defining the counting rules:

Rule 1: If any one of K mutually exclusive and exhaustive events can occur on each of N trials, then there are K_N different sequences that may result from a set of trials.

Rule 2: If K_1, …, K_N are the numbers of distinct events that can occur on trials 1, …, N in a series, then the number of different sequences of N events that can occur is $(K_1)(K_2)$ … (K_N).

Rule 3: The number of different ways that N distinct things may be arranged in order is N! = (1)(2)(3) … (N – 1)(N), where 0! = 1. An arrangement in order is called permutation, so that the total number of permutations of N objects is N! The symbol N! is called "N factorial."

A more generic notation is:

$$n! \equiv (n)\ (n-1)\ (n-2)...\ (n-(n-2))\ (2)\ (1) = n(n-1)!$$

where 1 ! = 1 and 0! ≡ 1 but (–n)! is undefined.

Stirling's approximation to n!: For large values of n it is difficult to determine the value of n! Therefore, we can use the approximation:

$$n! \approx \sqrt{2\Pi n}\quad n^n e^{-n}$$

Note: Most computers, including hand-held calculators, have a factorial key n! which reduces the need to apply Stirling's approximation.

Rule 4: The number of ways of selecting and arranging r objects from among N distinct objects is N!/(N – r)!

Rule 5: The total number of ways of selecting r distinct combinations of N objects, irrespective of order, is $N!/r!(N-r) = \binom{N}{r}$

PERMUTATIONS AND COMBINATIONS

SAMPLING

1. Sampling *without* replacement
   - Used to arrange distinct objects in some order.
   - Permutations: each ordering of distinct objects is unique.
   - Combinations: ordering of objects is irrelevant.
   - Number of permutations is greater than number of combinations.
   - Helps determine probability of "equally likely" outcomes.
   - Appears in definition of discrete hypergeometric distribution.

EXAMPLE

Consider a population of three letters: {a, b, c} = 3 = n

Experiment: Draw two letters: (x, y) = 2 = r

Sampling without replacement: $n \cdot (n-1) \cdot (n-2) \ldots (n-(r-1))$

Can form six "sample sets": $6 = 3 \cdot 2 = n \cdot (n-1)$

(a, b) (a, c) (b, c) (b, a) (c, a) (c, b)

2. Sampling *with* replacement

Can form nine "sample sets": $9 = 3^2 = n^r$

(a, a) (a, b) (a, c) (b, b) (b, c) (b, a) (c, c) (c, a) (c, b)

PERMUTATIONS

1. Each Ordering Is Unique

A permutation is a particular sequence of objects (or simple events) where the order of selection forms subsets or arrangements that are considered unique or distinctive.

For example, the three letters (abc) are considered unique and distinct from the other five possible arrangements of these letters:

$$(abc) \neq (acb) \neq (bac) \neq (bca) \neq (cab) \neq (cba)$$

Therefore, given n distinct objects arrange r of them in a set.
Assume:

a. Each sample is "equally likely."
b. Each sample is taken "without replacement."

Permutation of n objects taken r at a time:

$$P(r;n) = (n)(n-1) \ \ldots \ (n-(r-1))$$

$$= \frac{(n)(n-1) \ \ldots \ (n-(r-1)) \cdot (n-r)(n-(r+1)) \ \ldots \ 1)}{(n-r)(n-(r+1)) \ \ldots \ 1}$$

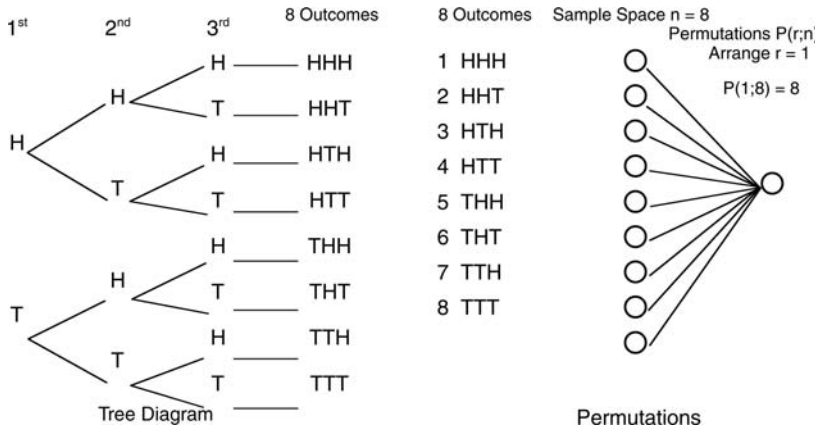
$$= \frac{n!}{(n-r)!}$$

Remember that by definition $0! \equiv 1$

Example

Single toss of three coins

Eight distinct outcomes: $n = 2 \times 2 \times 2 = 8$



Permutation of n objects taken r at a time:

$$P(r;n) = (n)(n-1) \ \ldots \ (n-(r-1))$$

$$= \frac{n!}{(n-r)!}$$

Also written as $P(n,r)$, ${}_nP_r$, $P_{n,r}$, P_r^n

Permutations of EIGHT objects taken ONE at a time:

$$P(1;8) = (8)(8-1) \ \ldots \ (8-(1-1)) = 8$$

$$= \frac{n!}{(n-r)!} = \frac{8!}{(8-1)!} = \frac{8!}{7!} = 8$$

2. Permutations Are Choosing "Without Replacement"

8 Objects	7 Remain	6 Remain	5 Remain
○	○	○	○
○	○	○	○
○	○	○	○
○	○	○	○
○	○	○	○
○	○	○	
○	○		
○			

a. Permutations: P(r;n)

r = 1 ○ → P(1;8) = 8

r = 2 ○○ → P(2;8) = 8 × 7 = 56

r = 3 ○○○ → P(3;8) = 8 × 7 × 6 = 336

b. Permutations of eight objects taken two at a time:

$$P(2;8) = (8)(8-1) \ \ldots \ (8-(2-1)) = (8)(7) = 56$$

$$= \frac{n!}{(n-r)!} = \frac{8!}{(8-2)!} = \frac{8!}{6!} = 8 \cdot 7 = 56$$

c. Permutations of eight objects taken three at a time:

$$P(3;8) = (8)(8-1) \ \ldots \ (8-(3-1)) = (8)(7)(6) = 336$$

$$= \frac{n!}{(n-r)!} = \frac{8!}{(8-3)!} = \frac{8!}{5!} = 336$$

3. Permutations of Different Types of Objects

If n objects are comprised of k unique types such that:

$$n = n_1 + n_2 + \ldots + n_k$$

Then the number of different permutations of n objects taken n at a time is:

$$P(n_1, n_2, \ldots n_k; n) = \frac{n!}{n_1! n_2! \ldots n_k!}$$

Example 1

Eight balls in an urn comprised of the following:

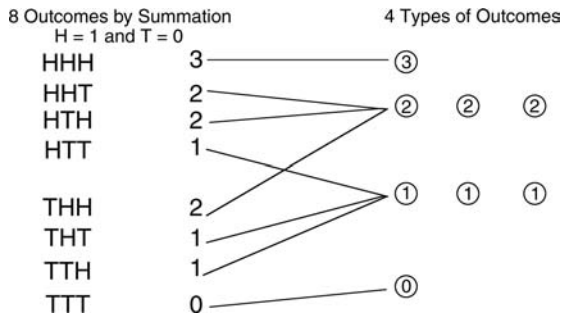
$$n_1 = 5 \text{ red}; \quad n_2 = 3 \text{ white}; \quad n_3 = 0 \text{ black}$$

$$P(5, 3, 0; 8) = \frac{8!}{5!3!0!} = \frac{8 \cdot 7 \cdot 6}{(3 \cdot 2 \cdot 1) \cdot 1} = 56$$

Example 2

Single toss of three coins

If we do not distinguish order of heads and tails in each of the eight total outcomes



Four Types of Objects Regardless of Order of H's and T's		Total n = 8 $P(n_1, n_2, n_3, n_4; n)$
		P(1, 3, 3, 1;8)
③	$n_1 = 1$	
② ② ②	$n_2 = 3$	$= \frac{8!}{1!3!3!1!}$
① ① ①	$n_3 = 3$	
⓪	$n_4 = 1$	= 1120

$$n = n_1 + n_2 + n_3 + n_4$$

$$P(n_1, n_2, \ldots n_k; n) = \frac{n!}{n_1! n_2! \ldots n_k!}$$

$$P(1, 3, 3, 1; 8) = \frac{8!}{1!3!3!1!} = 1120$$

Combinations — Ordering Is Irrelevant

A combination is a particular sequence of objects (or simple events) selected to form subsets or arrangements *without regard to the order* of objects in the arrangement. For example, the three letters (abc) are considered indistinguishable from these five other possible arrangements of these letters:

$$(abc) = (acb) = (bac) = (bca) = (cab) = (cba)$$

Therefore, given n distinct objects arrange r of them in a set.
Combination of n objects taken r at a time; ($r \leq n$):
Assume:
Each sample is "equally likely"

a. Each sample taken "without replacement"

$$C(r;n) \equiv \binom{n}{r} = \frac{(n)(n-1)\ \ldots\ (n-(r-1))}{r!}$$

$$= \frac{n!}{r!(n-r)!} = \frac{P(n,r)}{r!}$$

Note: The number of combinations is less than the number of permutations. The reader should also note that the combination formula may be written in two different ways as:

(1)
$$C(r;n) \equiv \binom{n}{r} = \frac{(n)(n-1)\ \ldots\ (n-(r-1))}{r!}$$

$$= \frac{n!}{r!(n-r)!} = \frac{P(r;n)}{r!}$$

and (2) $C(n,r),\ {}_nC_r,\ C_{n,r},\ C_r^n$

Permutation or Combination?

Example 1

How many ways can a group of eight people be chosen to form a committee of five members?

Assume a full five-member committee is formed. Each person is unique and selected "without replacement." Also, no distinction is made in terms of the order of selection of the "people." Hence, abcde = bcdea implies a combination.

Combination of 8 people (objects) taken 5 at a time:

$$C(5;8) = \frac{(8)(8-1)\ \ldots\ (8-(5-1))}{5!}$$

$$= \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{8 \cdot 7}{1} = 56$$

Example 2

How many ways can a group of eight people be seated at a head table having only five chairs?

Assume all five chairs are filled. Each person is unique and selected "without replacement."

The order of seating is important and distinguishes the seating plan, e.g., the seating order abcde ≠ bcdea. This implies a permutation and not a combination.

So, permutation of 8 people (objects) taken 5 at a time:

$$P(5;8) = (8)(8-1) \ldots (8-(5-1))$$
$$= 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 = 6720$$

Example 3

How many different ways can a group of five candidates be selected when (1) at least one award total will be presented, AND (2) each candidate can receive at most only one award?

Assume at least one award is made.

Assume one person cannot receive two awards. (i.e., sampling "without replacement") All awards are equal; order of selection not important. Therefore, the problem is of combination not permutation in nature.

Solution:

One out of 5 could be selected:	$C(1;5) = 5$
Two out of 5 could be selected:	$C(2;5) = 10$
Three out of 5 could be selected:	$C(3;5) = 10$
Four out of 5 could be selected:	$C(4;5) = 5$
Five out of 5 could be selected:	$C(5;5) = 1$

Total number of possible awards that could be presented:

$$C(1;5) + C(2;5) + C(3;5) + C(4;5) + C(5;5)$$
$$= 5 + 10 + 10 + 5 + 1 = 31$$

Binomial Expansion

As we already have mentioned, permutations, combinations and factorials are used in the binomial distributions. Let us see how.

Combinations

Combinations are also called "binomial coefficients," which arise in the expression for a Binomial Expansion. To identify this expansion given n objects and arranging r of them in a set, we use the following notation:

$$C(r;n) \equiv \binom{n}{r} = \frac{(n)(n-1)\ \ldots\ (n-(r-1))}{r!}$$

$$= \frac{n!}{r!(n-r)!}; \quad r \le n$$

Properties of Binomial Coefficients

Since: n – (n – r) = r then a symmetry exists:

$$\binom{n}{n-r} = \binom{n}{r}$$

that is, if r + k = n

then $$\binom{n}{k} = \binom{n}{r}$$

can show: $$\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}$$

Note: $\binom{a}{b} \equiv 0$ if b > a Since (−a)! Undefined

Binomial Expansion

Fundamentally, the binomial expansion is a power function and is based on the successive powers of the given function of the form

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$

$$\sum_{k=0}^{n} \frac{n!}{k!(n-k)!}\ a^{n-k} b^k$$

Examples of this expansion for n = 0 up to n = 5 are given below.

$n = 0$: $(a + b)^0 = 1$

$n = 1$: $(a + b)^1 = a + b$

$n = 2$: $(a + b)^2 = a^2 + 2ab + b^2$

$n = 3$: $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$

$n = 4$: $(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$

$n = 5$: $(a + b)^5 = a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$

We can also use the same rationale to use the expansion for other useful factors and expansions, such as in the following examples:

$$a^2 - b^2 = (a + b)(a - b)$$

$$a^2 + b^2 = (a + b)(a - b)$$

$$a^3 + b^3 = (a + b)(a^2 - ab + b^2)$$

$$a^3 - b^3 = (a - b)(a^2 + ab + b^2)$$

$$a^4 + b^4 = (a^2 + ab\sqrt{2} + b^2)(a^2 - ab\sqrt{2} + b^2)$$

As a quick way to figure out the expansion, one may use the Pascal's Triangle for the $(a + b)^n$ array of coefficients of successive powers of $(a + b)^n$.

n																	
0									**1**								
1								**1**		**1**							
2							**1**		**2**		**1**						
3						**1**		**3**		**3**		**1**					
4					**1**		**4**		**6**		**4**		**1**				
5				**1**		**5**		**10**		**10**		**5**		**1**			
6			**1**		**6**		**15**		**20**		**15**		**6**		**1**		
7																	
8																	

The reader should try to fill the coefficients for n = 7 and 8. Some general guidelines to generate the coefficients are:

1. A power of n has (n + 1) terms.
   If n is odd the number of terms is even.
   If n is even the number of terms is odd.
2. The first and last numbers in each row are always 1.
3. The second number and the next-to-last number are equal to n.
4. The numerical values of the remaining inner terms can be determined by addition of the two numbers appearing directly above, e.g., for row n = 6, the third term is 15 which is the sum of the 10 and 5 appearing above its position in row n = 5. The fourth is 20 which is the sum of 10 and 10 appearing above its position in row n = 5. The fifth is 15 using the same rationale and so on.

16 Discrete and Continuous Random Variables

A special terminology is useful for discussing probability distributions of numerical scores. In this chapter we will discuss discrete and continuous random variables with some associated distributions. Some of the material in this chapter is used with permission from Stuart, A., Ford Motor Company.

INTRODUCTION

A random variable (RV) is a real valued number assigned to each individual sample, S_i, according to some function $X(S_i)$

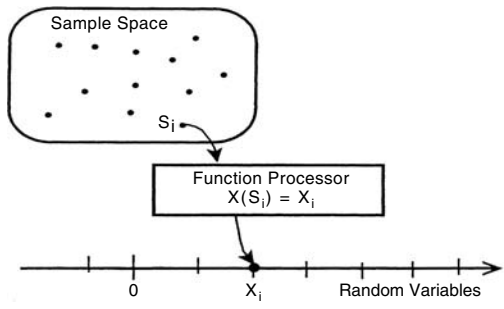
Individual sample, S_i
Discrete real-valued number, X_i.
An individual sample cannot be assigned two random values.
Range of real-numbers is typically finite.

$$x = \{X_i\} \qquad i = 1, 2, 3, \ldots, M$$

or

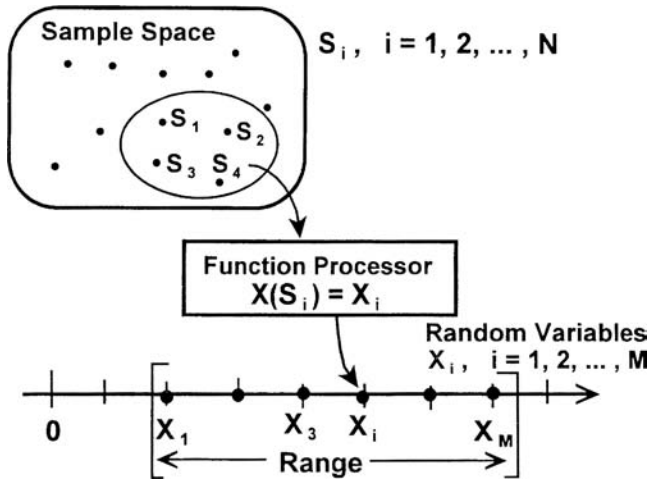
$$x = \{X_1, X_2, X_3, \ldots, X_M\}$$

Assume an ascending order of magnitude: $X_{i-1} < X_i < X_{i+1}$. Pictorially, this may be shown as:



SAMPLES ASSIGNED THE SAME RANDOM VARIABLE

1. Two or more individual samples may be assigned the same numerical random value X_i.
2. The total number of random variables M will be less than or equal to the total number of individual samples N.
3. Range of random variables: $[X_1 \le X_i \le X_M]$
   (There are M discrete random variables in the range.)
4. If all samples are assigned a unique RV, then M = N. This may be shown in a pictorial form as:



RANDOM VARIABLES GROUPED INTO CELLS

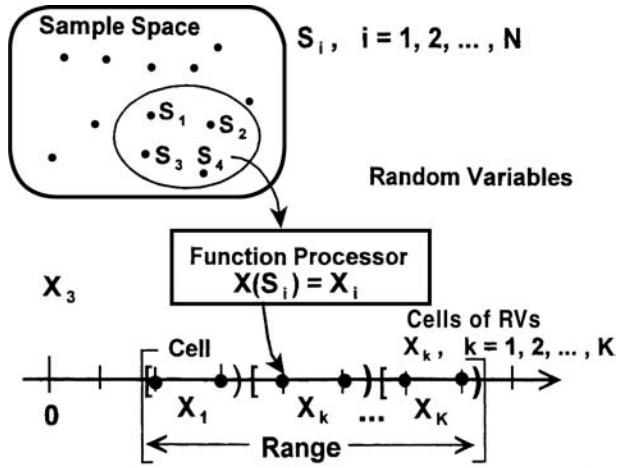
1. Group random variables into cells (or "bins") if the number of random variables is large, say M > 30.
2. Groups or cells are formed in the numerical RV domain by dividing the range into say 10 to 20 equal cells.
3. While the total number of RVs is M, the number of cells is K.
4. All the individual RVs in a cell assume the value of the RV in, say the center of the cell interval and are denoted X_k.
5. Each cell or grouping is designated X_k and represents a finite interval of contiguous RVs.

Individual: $X_i = [(X_1, X_2, \ldots), \ldots, [X_{i-1}, \ldots, (X_{M-1}, X_M)]$

Cells: $X_k = [(X_1), \ldots, (X_k), \ldots, [X_K)]$

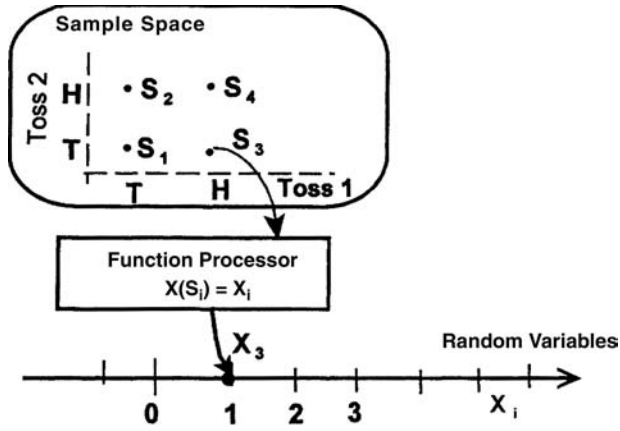
where $X_k = X_{i-1} \le X_i < X_{ik}$

This may be shown as:



Random Experiment

Two tosses of a coin.



Sample Space: $S_1 = TT$, $S_2 = TH$, $S_3 = HT$, $S_4 = HH$
Function Process:

1. Assign real-value number to occurrence: T = 0 and H = 1.
2. Sum of real-values for individual samples S_i is RV X_i.

$$S_1 = TT \rightarrow X_1 = 0 + 0 = 0$$

$$S_2 = TH \rightarrow X_2 = 0 + 1 = 1$$

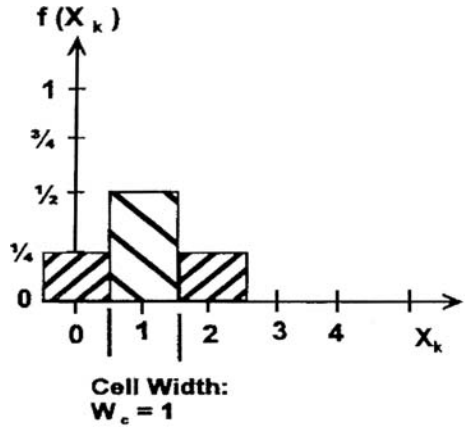


FIGURE 16.1 Discrete probability density function.

$$S_3 = HT \rightarrow X_3 = 1 + 0 = 1$$

$$S_4 = HH \rightarrow X_4 = 1 + 1 = 2$$

Random Variables: arranged in ascending order by magnitude

$$\{S_i\} = \{TT, TH, HT, HH\} \rightarrow \{X_i\} = \{0, 1, 1, 2\}$$

Note: cell width $W_c = 1$.

DISCRETE PROBABILITY DISTRIBUTION

Discrete probability density function: probability assigned to AREA of discrete random variable cells.

$$\text{Probability (Cell area)} = \text{Height } (X_k) \cdot \text{Width } (W_c)$$

Discrete probability is the frequency of the grouped values $x = X_k$ for the data corresponding to the individual event S_i in sample space. This may be shown in a pictorial form in Figure 16.1.

However, when data are grouped as consecutive integer values the cell width is unity; $W_c = 1$.

1. Probability of group data with cell width W_c:

$$P(x = X_k) = f(X_k) \cdot W_c;\ k = 1, 2, 3, \ldots, K$$

where $f(X_k)$ is the probability density for the RV cell X_k, which has a sample frequency f_k for a total sample size of n:

$$f(X_k) \equiv f_k/n$$

2. Discrete probability density function properties:
   a. Positive $0 < P(X_k) = f(X_k) \cdot W_c$
   b. Unit area (area is sum under curve unity)

$$\sum_{k=1}^{K} f(X_k) \cdot W_c = 1$$

Random Experiment

Two tosses of a coin.

Random variable of sample event X_i is defined as the number of heads to appear in two tosses.

$$\{S_i\} = \{TT, TH, HT, HH\} \rightarrow \{X_i\} = \{0, 1, 1, 2\}$$

SAMPLE SPACE: TT TH HT HH
Probability (sample space): 1/4 1/4 1/4 1/4
Random Variable X_i: 0 1 1 2 $\rightarrow$ ascending (Number of Heads)
Grouping into four cells each of width, $W_k = 1$;
Discrete probability density function $f(X_k)$:

$$f(0) = 1/4; \quad f(1) = 1/4 + 1/4 = 1/2; \quad f(2) = 1/4$$

The above may be represented in graphic displays as in Figure 16.2.

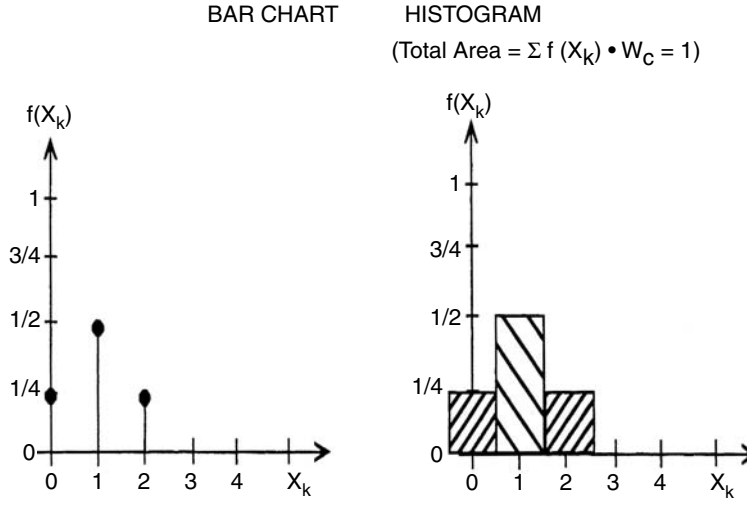


FIGURE 16.2 A bar chart and a histogram of two tosses of a coin.

DISCRETE CUMULATIVE DISTRIBUTION FUNCTION

Probability that value of the random variables X_k will be less than or equal to a specified value X_m.

$$F(X_m) = P(X_k \le X_m) = \sum_{k=1}^{m} P(X_k) = \sum_{k=1}^{m} f(X_k) \cdot W_c$$

where the capital letter F is used for cumulative distribution. If sample space is a finite number K of random variables:

$$x = \{X_1, X_2, X_3, \ldots, X_k, \ldots, X_m, \ldots, X_K\}$$

where we assume an ascending order: $X_{k-1} < X_k < X_{k+1}$.

For example, if m = 2:

$$\begin{aligned} F(X_2) &= f(X_1) \cdot W_c + f(X_2) \cdot W_c \\ &= P(X_1) + P(X_2) \\ &= P(X_k \le X_2) \end{aligned}$$

Upper bound: m = K

$$F(X_K) = P(X_k \le X_K) = 1$$

Random Experiment

Two tosses of a coin.

Random Variable X_k defined as number of heads. (Grouped data X_k has a cell width of unity $W_c = 1$.)

R.V.X_k	Prob. Fn. $f(X_k)$	Cumulative Prob. $F(X_k)$
$X_k < 0$	$P(X < X_1) = 0$	
$X_1 = 0$	$f(X_1) = 1/4$	$F(X_1) = P(x \le X_1) = 1/4$
$X_2 = 1$	$f(X_2) = 1/2$	$F(X_2) = P(x \le X_2)$
		$= P(X_1) + P(X_2)$
		$= 1/4 + 1/2 = 3/4$
$X_3 = 2$	$f(X_3) = 1/4$	$F(X_3) = P(x \le X_3)$
		$= P(X_1) + P(X_2) + P(X_3)$
		$= 1/4 + 1/2 + 1/4 = 1$

Figure 16.3 shows the cumulative distribution of two tosses of a coin.

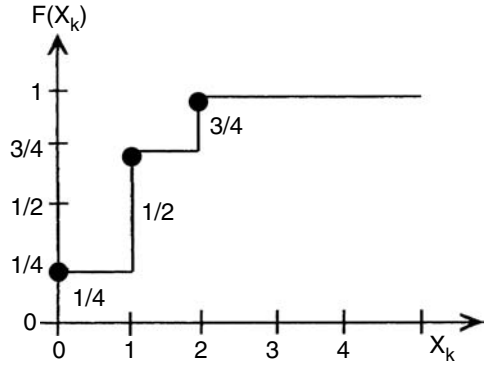
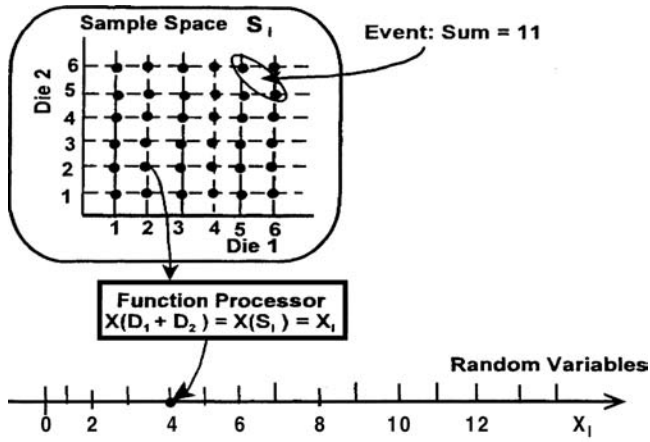


FIGURE 16.3 Cumulative distribution of two tosses of a coin.

Random Experiment

Toss a pair of fair dice.



Individual Sample S_i: Defined as sum of face values on pair of six-sided dice.

$$S_i = D_1 + D_2$$

Total sample space size: $N = 6 \times 6 = 36$ possible outcomes
Random Variable X_k: Is cell equal to a specific value of S_i

$$X_{k\,=}\{S_i\} = \{D_1 + D_2\}$$

It is quite common in engineering to have numerical value samples.
Total number of Random Variables cells: $M = 11$
Range of RV: $[X_1 = 2 \leq X_k \leq 12 = X_{11}]$

Other possible processes (and distributions) could include:

1. Product: $S_i = D_1 \cdot D_2$
2. Magnitude of difference: $S_i = |D_1 - D_2|$
3. Ratio: $S_i = D_1/D_2$

Random Variable: $(X_k = \{S_i\} = \{D_1 + D_2\})$

$$X \text{ (Die 1 + Die 2)} = X(\text{Sum}) = \text{Sum} = X_{(\text{sum}-1)}$$

Example outcome:

$$X \text{ ([Die 1 = 1] + [Die 2 = 2])} = X(1 + 2) = 3 = X_2$$

Consider an event A: Set of all RVs equal to 3

$$\{X_2\} = \{[1 + 2], [2 + 1]\} = \{3, 3\}$$

Sample size of event A is therefore m = 2
Probability Density: particular event $\{X_2\}$

$$f(X_2) = P(X_2) = m/N$$

or

$$f(3) = P(3) = 2/36$$

Cumulative distribution: for random variable $X_2 = 3$

$$F(X_2) = f(X_1) + f(X_2)$$

or

$$\begin{aligned} F(3) &= f(2) + f(3) \\ &= 1/36 + 2/36 = 3/36 \end{aligned}$$

$S_i = D_1 + D_2$ = Sum of numbers appearing on face cell: $X_k = \{S_i\} = \{D_1 + D_2\}$

The probabilities associated with each cell are shown in Table 16.1
Figure 16.4 shows the probability density function and Figure 16.5 shows the cumulative probability function.

TABLE 16.1
Probability Density and Distribution of a Pair of Fair Dice

R.V. X_k	No. Outcomes with Value X_k	$f(X_k)$	$F(X_k)$
2	1	1/36	1/36
3	2	2/36	3/36
4	3	3/36	6/36
5	4	4/36	10/36
6	5	5/36	15/36
7	6	6/36	21/36
8	5	5/36	26/36
9	4	4/36	30/36
10	3	3/36	33/36
11	2	2/36	35/36
12	1	1/36	36/36
	Sum = 36	36/36	

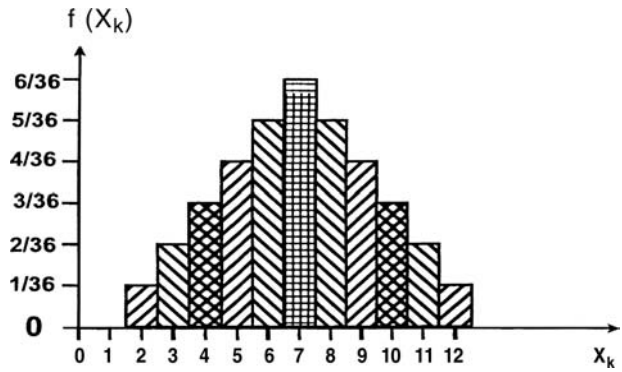


FIGURE 16.4 Probability density function.

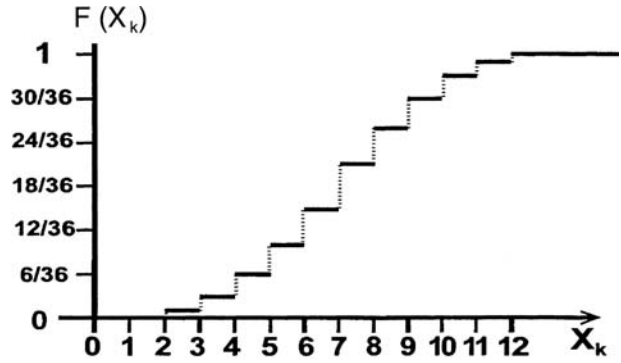


FIGURE 16.5 Cumulative probability function.

MEAN OR EXPECTED VALUE

MEAN from RV Range: M discrete values or cells, X_k

$$\overline{X} \equiv E[X] = \sum_{k=1}^{M} x_k f(X_k)$$

or MEAN from Sample Space: N discrete individual samples, S_i

$$\overline{X} = 1/N \sum_{i=1}^{N} X(S_i)$$

Random Experiment

Sum produced by pair of fair six-sided dice.

Random Variable X_k defined as sum of the two numbers:

$$X_k = \{D_1 + D_2\}$$

Cell k	No. in X_k	X_k	f (X_k)	X_k f (X_k)
1	1	2	1/36	2/36
2	2	3	2/36	6/36
3	3	4	3/36	12/36
4	4	5	4/36	20/36
5	5	6	5/36	30/36
6	6	7	6/36	42/36
7	5	8	5/36	40/36
8	4	9	4/36	36/36
9	3	10	3/36	30/36
10	2	11	2/36	22/36
11	1	12	1/36	12/36
M = 11	36			Sum 252/36

Mean $$\overline{X} = \sum_{k=1}^{11} X_k f(X_k) = 252/36 = 7$$

Sample variance and standard deviation

1. Sample variance: Expected value of X_i about the mean

$$s_x^2 = \sum_{k=1}^{M} (X_k - \overline{X})^2 f(X_k)$$

or

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X(S_1)-\bar{X})^2$$

2. Sample standard deviation: $s_x \equiv \sqrt{E[(x-\bar{x})^2]}$

Exercise: Sum of two fair dice: $X_k = \{D_1 + D_2\}$

x_k	$(X_k - \bar{X})$	$(X_k - \bar{X})^2$	$f(X_k)$	$(X_k - \bar{X})^2 f(X_k)$
2	–5	25	1/36	25/36
3	–4	16	2/36	32/36
4	–3	9	3/36	27/36
5	–2	4	4/36	16/36
6	–1	1	5/36	5/36
7	0	0	6/36	0
8	1	1	5/36	5/36
9	2	4	4/36	16/36
10	3	9	3/36	27/36
11	4	16	2/36	32/36
12	5	25	1/36	25/36
				Sum 210/36

Variance:

$$s_x^2 = \sum_{k=1}^{11}(x_k-\bar{x})^2 f(x_k) = \frac{210}{36} = 5\frac{5}{6} = 5.833$$

Standard deviation:

$$s_x = 2.415$$

CONTINUOUS RANDOM VARIABLES

As we have mentioned earlier, discrete random variables are represented by isolated real-valued numbers. Manipulation of discrete random variables involves summations of these discrete values. This is illustrated for the frequency of outcomes in k-cells; n = total number samples.

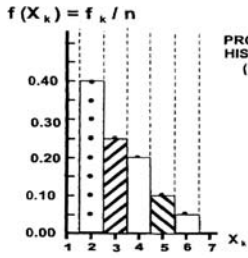
Probability density function (pdf): $f(X_k) = \frac{f_k}{n}$

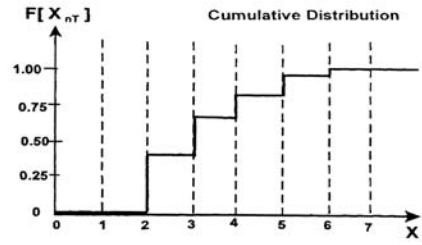
Cumulative distribution function (CDF) (sum)

$$F[X_{nT}] = \sum_{k=1}^{n_T} f(X_k)$$

Mean $$\overline{X} = E[X_k] = \sum_{k=1}^{n_c} X_k f(X_k)$$

Variance $$S^2 = E[(X_k - \overline{X})^2] = \sum_{k=1}^{n_c} (X_k - \overline{X})^2 f(X_k)$$





Advantages of Continuous Random Variables

- Integrals (areas) of continuous r.v. x yield closed form equations that are easy to manipulate and to analyze.
- Integrals of standardized distributions can be tabulated.

Probability density function (pdf): f (x) where probabilities are only defined as the area within an interval.

$$P(a \le X \le b) = \int_a^b f(x)dx$$

Two constraints of a probability density function: f (x)

1. Positive value: $f(x) \ge 0$
2. Unit area: $\int_{-\infty}^{x} f(x)dx \equiv 1$

Cumulative distribution function (CDF): $F(x) = (P(-\infty < X \le x) = \int_{-\infty}^{x} f(x)dx$

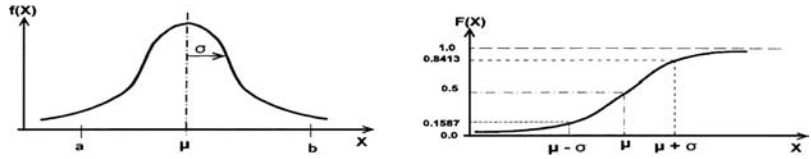


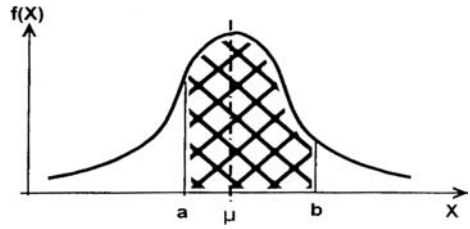
FIGURE 16.6 The probability (left) and cumulative (right) functions.

$$f(x) = dF(x)/dx$$

These functions may be represented by the graphs in Figure 16.6.

Properties of Continuous Distributions

1. Probability is defined only in the context of an incremental range of the random variable [$a \le x \le b$].



$$P(a \le X \le b) = \int_a^b f(x)dx = \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx = F(b) - F(a)$$

2. Probabilities cannot be determined for a point x_0, since the interval of integration or the base (b – a) is zero.
3. Probabilities can be determined from either the probability density function f(x) or the cumulative distribution function F(x).
4. The CDF is more important because tabulated values of the various standardized or normalized probability distributions models presented in this form.
5. Mean:

$$\mu_x = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

6. Variance:

$$\sigma_x^2 \equiv E[(X - \mu_x)^2] = \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x)dx = E(X^2) - \mu_x^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu_x^2$$

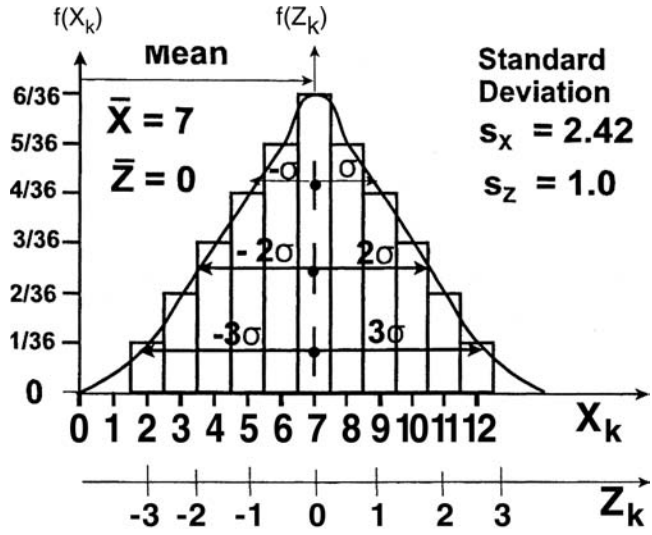


FIGURE 16.7 The normal distribution.

STANDARDIZED RANDOM VARIABLE

Sampling histograms that are symmetric and without outliers can be converted to standardized random variables:

$$Z_k \equiv \frac{X_k - \overline{X}}{s}$$

which has (1) mean of r.v.Z_k at the origin: Zbar = 0; and (2) variance of r.v. Z_k of any unity: $S_z^2 = 1$.

This standardized form of the sample r.v. conveniently presents results in terms of the number of standard deviations above or below the mean. This number is often called the "Z-score" and appears as:

$$P(\overline{X} - Z \cdot s_x \le X_k \le \overline{X} + Z \cdot s_x)$$

A typical pictorial view of this type of distribution is shown in Figure 16.7.

TYPICAL UNSTANDARDIZED FORM OF TABULATED CDF

Leading Tail Interval

$$(-\infty < x \le b)\{where \quad F(-\infty) \equiv 0\}$$

This may be shown mathematically as

$$P(-\infty < x \le b) \equiv P(x \le b) = \int_{-\infty}^{b} f(x)dx = F(b) - F(-\infty) = F(b)$$

Trailing Tail Interval

$$(b \le x < \infty) \text{ where } F(\infty) \equiv 1$$

This is shown mathematically as

$$P(b \le X < \infty) \equiv P(X \ge b) = \int_{b}^{\infty} f(x)dx = F(\infty) - F(b) = 1 - F(b)$$

Upper Range from Mean Value

$$(\mu \le X \le b)\left\{ where \quad F(\mu) \equiv \frac{1}{2} \right\}$$

This may be shown mathematically as

$$P(\mu \le X \le b) = \int_{\mu}^{b} f(x)dx = F(b) - F(\mu)$$

PROBABILITY DISTRIBUTION

We already have seen that this distribution follows the shape as in Figure 16.7. However, in addition to the shape one may determine confidence intervals. These confidence intervals represent the Cumulative Probability Within Interval about Mean, and they may be represented mathematically (assuming normality) as:

$$\begin{aligned} &P\ (\overline{X} - Z \cdot s_x \le X_k \le \overline{X} + Z \cdot s_x) \\ &= F(X = (\overline{X} + Z \cdot s_x)) - F(X = (\overline{X} - Z \cdot s_x)) \end{aligned}$$

Z	Interval	F(Interval)
0.675	$\overline{X} \le \overline{X}_x \le X + 0.675s_x + 0.675s_x$	0.500
1	$\overline{X} - 1S_x \le X_k \le \overline{X} + 1S_x$	0.6827
2	$\overline{X} - 2S_x \le X_k \le \overline{X} + 2S_x$	0.9545
3	$\overline{X} - 3S_x \le X_k \le \overline{X} + 3S_x$	0.9973

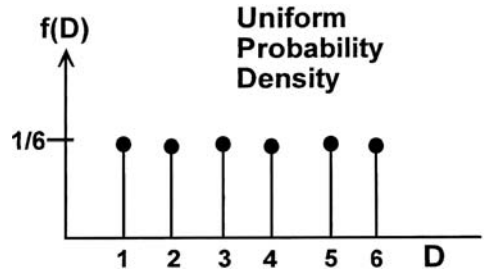


FIGURE 16.8 Uniform probability density for a die.

UNIFORM DISTRIBUTION

If we consider each die as an independent random variable, then a uniform probability distribution will be the result and will take the shape shown in Figure 16.8.

With mean value of single die:

$$\overline{D} = \frac{1}{6}[1+2+3+4+5+6] = \frac{1}{6}[21] = 3.5$$

Variance about mean:

$$S_D^2 = \frac{1}{6}[(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2]$$

$$= \frac{1}{6}[(-2.5)^2 + (-1.5)^2 + (-0.5)^2 + (0.5)^2 + (1.5)^2 + (2.5)^2]$$

$$= \frac{2}{6}[(0.5)^2 + (1.5)^2 + (2.5)^2] = \frac{2}{6}[8.75] = 2.9166$$

Standard deviation: $s_D = 1.708$

The uniform distribution may be represented in general mathematical notation along with the constraints as:

$$F(X;a,b) = \begin{cases} 0 & x < a \\ \dfrac{1}{b-a} & a \le x \le b \\ 0 & x > b \end{cases}$$

and the shape looks like Figure 16.9.

When checking for the probability density function of a uniform distribution one should check the following two properties:

1. Positive: f {X; a,b) ≥ O

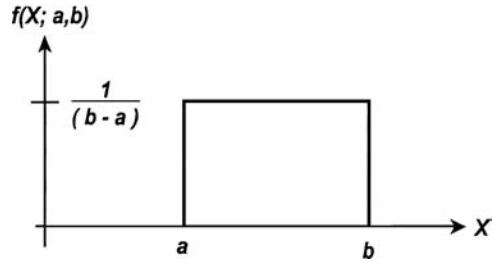


FIGURE 16.9 A generic uniform distribution.

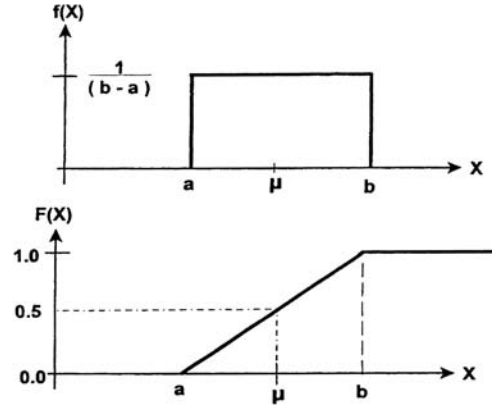


FIGURE 16.10 A comparison of the uniform distribution and its C.D.F.

2. Unit area:

$$\int_a^b \frac{1}{b-a}\,dx = \frac{1}{b-a}\,x\Big|_a^b = \frac{1}{b-a}(b-a) = 1$$

Note: Height of p.d.f. is equal to the reciprocal of the base. Figure 16.10 shows the comparison of the uniform distribution and the cumulative density function.

The mathematical notation of the c.d.f. with the constraints is:

$$\mathrm{F(x)} \equiv \mathrm{P}(-\infty < X \le x) = \int_{-\infty}^{x} \frac{1}{(b-a)}[U(x-a) - U(x-b)]dx$$

$$\mathrm{F(X)} \equiv P(-\infty < x \le x) \equiv \begin{cases} 0 & x < a \\ \dfrac{x-a}{(b-a)} & a \le x \le b \\ 1 & b \le x \end{cases}$$

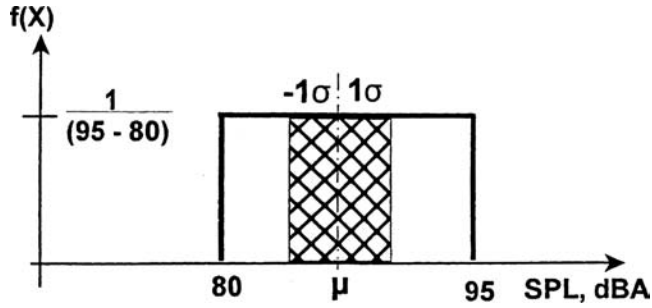


FIGURE 16.11 The sound level in a room.

Mean: $$\mu_x = \frac{b+a}{2}$$

Variance: $$\sigma_x^2 = \frac{(b-a)^2}{12}$$

The reader will notice that in the uniform distribution, (1) median = mean, and (2) there is no mode.

EXAMPLE

Sound level in a room is found to be uniformly distributed between 80 and 95 dBA. Occupational Safety and Health Administration (OSHA) regulations set a maximum safe level of 90 dBA for an 8-hour workday. See Figure 16.11.

Find:

1. The probability density function, f(x) for this noise
2. The probability of exceeding the 90 dBA standard
3. The mean and standard deviation
4. The level range within ± 1 σ of the mean
5. The probability of being in this ± 1 σ about the mean

Solutions:

1. f(X) = 1/15; $80 \le x \le 95$
2. Probability of exceeding 90 dBA:

$$P(> 90) = 1 - P(x \le 90) = 1 - F(x = 90)$$

$$1-\int_{-\infty}^{90} f(x)dx = 1-\frac{1}{15}\int_{80}^{90} dx$$

$$= 1-\frac{1}{15}x\Big|_{x=80}^{x=90} = 1-\frac{1}{15}(90-80)$$

$$= 1-\frac{10}{15} = \frac{5}{15} = \frac{1}{3} = 0.3333$$

3. Mean: $\mu = \frac{1}{2}(b+a) = \frac{1}{2}(95+80) = 87.5dBa$

Standard Deviation: $\sigma_x^2 = \frac{(b-a)^2}{12}$

$$\sigma_x = \sqrt{\frac{(95-80)^2}{12}} = \sqrt{\frac{(15)^2}{12}} = 4.333dBa$$

4. One-sigma range about mean: $(\mu - 1\sigma) \le x \le (\mu + 1\sigma)$

$$83.2 \le x \le 91.8$$

5. Probability of SPL being in this one-sigma range

$$P(83.2 \le x \le 91.8) = \int_{83.2}^{91.8} \frac{1}{15} dx = \frac{91.8-83.2}{15} = \frac{8.63}{15} = 0.575$$

NORMAL DISTRIBUTION — OTHERWISE KNOWN AS THE "BELL CURVE"

Figure 16.12 shows a typical normal curve with a mean and a standard deviation.

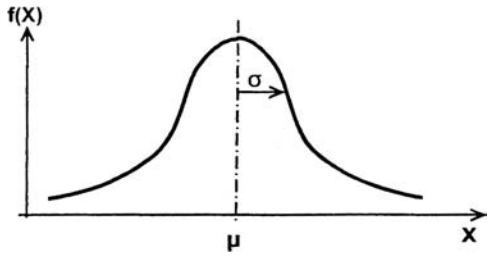


FIGURE 16.12 A typical normal curve.

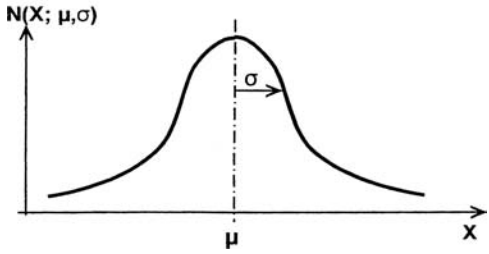


FIGURE 16.13 Probability density function for random variable x.

Typical Comments about the Normal Curve

- Most important distribution function in statistics.
- Many populations are normally distributed.
- Some distributions are made normal by changing variables.
- Normal can approximate discrete binomial distribution.
- Central Limit Theory → distribution of means is normal.
- Confidence limits are based on normal parameters.

Only two statistical parameters describe distribution:

μ = Mean
σ = Standard Deviation

Normal Probability Density Function is symmetric about mean.

Note: Mode = Mean = Median.

It turns out that the probability density function for variable x may be shown as in Figure 16.13.

Mathematically this is written as:

$$N(X = x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

with the following population parameters: μ = population mean, σ^2 = variance of population, σ = standard deviation of population, RV $-\infty < x < \infty$, inflection points at $x = \mu \pm \sigma$.

It is very important to realize that the area under the pdf is always unity even though the mean and standard deviation may differ. This is shown in Figures 16.13 and 16.14, respectively. Also in Figure 16.14 the probability density function is shown with different means and fixed standard deviation whereas the Figure 16.15 shows the probability density function with different means and/or standard deviations.

In addition, the normal curve may be represented as a cumulative distribution function. Its mathematical notation is

$$F(X) = \int_{-\infty}^{x} f(X)dX = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(X-\sigma)^2/2\sigma^2} dX$$

This is shown pictorially in Figure 16.16.

A word of caution — Any area of specified region under probability density function may be identified and calculated. However, the calculation cannot be a direct integration. Rather, because the normal probability function cannot be integrated directly, standardized cumulative distributions are tabulated. When those

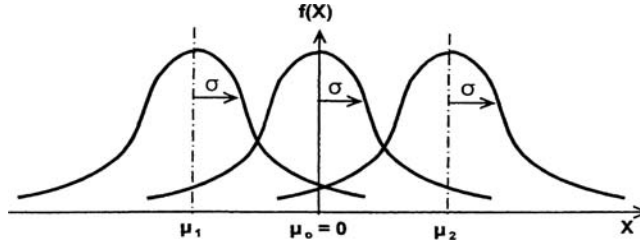


FIGURE 16.14 Probability density function with different means and same standard deviation.

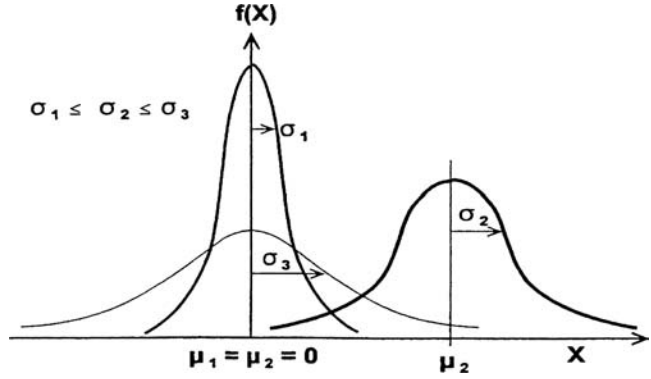


FIGURE 16.15 Probability density with different means and/or standard deviation.

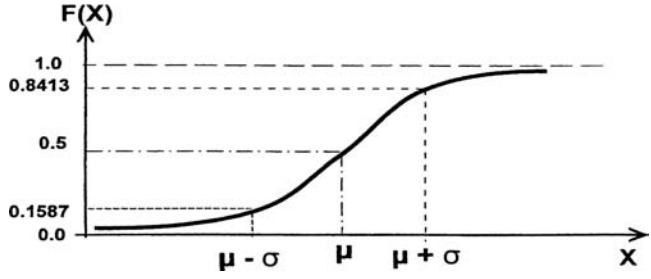


FIGURE 16.16 Cumulative distribution function.

tabulations are figured out, the reader must be very careful because not all tables give the same "area" of p.d.f. For example, some tables give S-CD values in interval: p ($0 \le Z \le Z_o$) and some tables give S-CD values of upper tail: P ($Z_o \le Z \le \infty$).

Differential Equation (DE)

The normal distribution functions can be generated from the first order homogeneous differential equation having a non-constant coefficient:

$$s^2 \frac{dY}{dX} + (X - \overline{X})Y = 0$$

where X is the independent random variable, Y is the dependent random variable (p.d.f.), $\overline{X}$ is the mean, and s^2 is the variance (units of X^2).

Separation of variables yields:

$$\frac{dY}{Y} = \frac{(X - \overline{X})}{s^2} dX$$

Solution of DE: $\ln Y = -(X - \overline{X})^2/(2s^2)$

Given the constraint that the area of the integral is unity, $\int_{-\infty}^{-\infty} Y\{X\}dX \equiv 1$. (The units of the probability density Y are the inverse of X.) It can be shown that the amplitude is adjusted so

$$Y(X) = \frac{1}{s\sqrt{2\pi}} e^{-(x-\overline{X})^2/2s^2}$$

Standardized Random Variables — Tabulated Function

Any random variable X can be transformed into a normalized (non-dimensional) standardized random variable Z

$$Z \equiv \frac{X - \mu_x}{\sigma_x} \quad \text{or} \quad Z \equiv \frac{X - \overline{X}}{s}$$

1. Subtracting the mean shifts mean of RV to origin: Z = 0.
2. Dividing by standard deviation makes variance unity and eliminates physical units.

This transformation is the function typically tabulated and plotted in statistics textbooks.

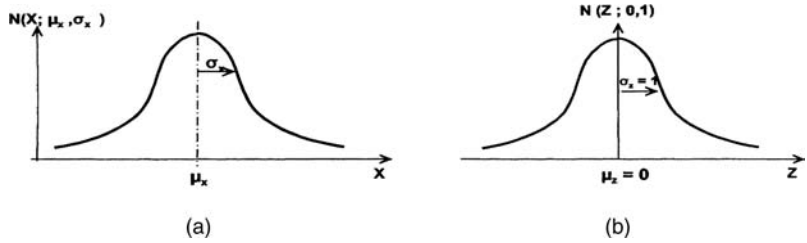


FIGURE 16.17 Standardized and unstandardized normal function: (a) unstandardized distribution, (b) standardized distribution.

Standardized Normal Distribution (SND)

Probability density function for the random variable Z:

$$f(Z = z) = N(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

where mean: $\mu_z = 0$, standard deviation: $\sigma_z = 1$, and RV $-\infty \le z \le \infty$. Pictorially this can be shown as in Figure 16.17. At the mean value, z = 0, the pdf is N (0; 0, 1) = 0.4. Also, this figure shows the unstandardized distribution as a comparison.

On the other hand, the cumulative distribution function (area under standardized probability density function for a given interval) of the standardized normal distribution may be presented mathematically as:

$$P(Z_1 \le Z \le Z_2) = \int_{z_1}^{z_2} f(Z)dZ = \int_{-\infty}^{z_2} f(Z)dZ - \int_{-\infty}^{z_1} f(Z)dZ$$

The graph is shown in Figure 16.18. Recall, however, that the probability for all continuous random variables can only be determined in terms of an interval such as $P(Z_1 \le Z \le Z_2)$.

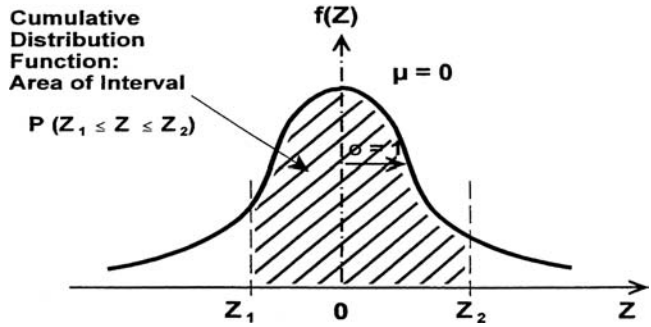


FIGURE 16.18 Cumulative distribution function — area of interval.

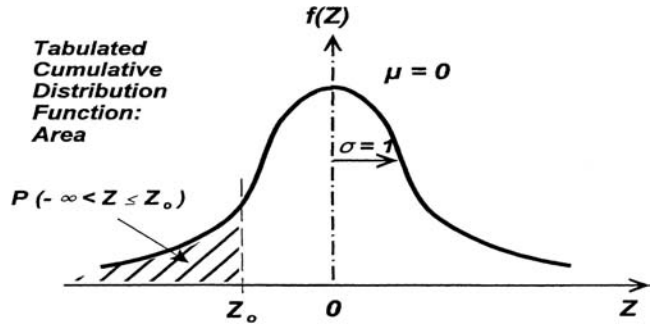


FIGURE 16.19 Tabulated cumulative distribution function — leading tail.

Tabulated possibilities include:

$$P(-\infty \le Z \le Z_0)$$
$$P(0 \le Z \le Z_0)$$
$$P(Z_0 \le Z \le \infty)$$

Figure 16.19 shows the tabulated distribution function for leading tail.

Tabulated Values (Leading Tail): $P(-\infty \le z \le z_0)$

$Z_0 = -3$	$p\ (-\infty \le Z \le -3) = 0.00135$
$Z_0 = -2$	$p\ (-\infty \le Z \le -2) = 0.0228$
$Z_0 = -1$	$p\ (-\infty \le Z \le -1) = 0.1587$
$Z_0 = 0$	$p\ (-\infty \le Z \le 0) = 0.5000$
$Z_0 = 1$	$p\ (-\infty \le Z \le 1) = 0.8413$
$Z_0 = 2$	$p\ (-\infty \le Z \le 2) = 0.9772$
$Z_0 = 3$	$p\ (-\infty \le Z \le 3) = 0.99865$

Figure 16.20 shows the tabulated cumulative distribution function under a specific interval area.

Tabulated Values (Center .Interval): $P(0 \le Z \le Z_0)$

$Z_0 = 0.0$:	$P(0 \le Z \le 0.0) = 0.0000$
$Z_0 = 0.5$:	$P(0 \le Z \le 0.5) = 0.1915$
$Z_0 = 1.0$:	$P(0 \le Z \le 1.0) = 0.3413$
$Z_0 = 1.5$:	$P(0 \le Z \le 1.5) = 0.4332$
$Z_0 = 2.0$:	$P(0 \le Z \le 2.0) = 0.4772$
$Z_0 = 2.5$:	$P(0 \le Z \le 2.5) = 0.4938$
$Z_0 = 3.0$:	$P(0 \le Z \le 3.0) = 0.4987$

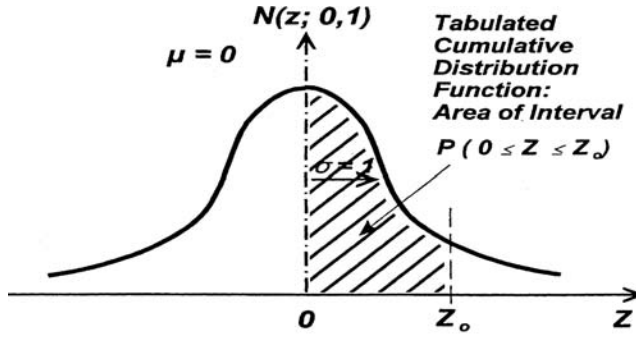


FIGURE 16.20 Tabulated cumulative distribution function — area of specific interval.

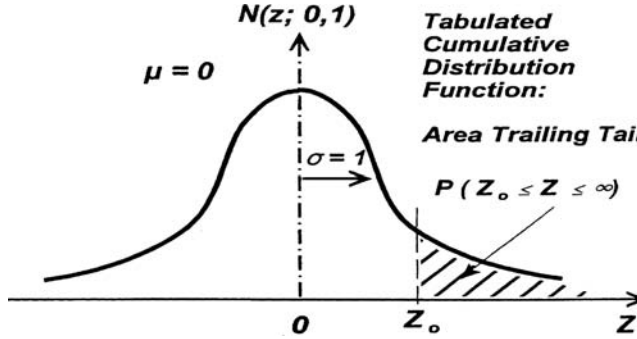


FIGURE 16.21 Standardized normal distribution with trailing tail.

Finally, the cumulative distribution function with a trailing tail is shown in Figure 16.21.

Tabulated Values (Trailing Tail): $P(Z_0 \leq Z \leq \infty)$

$Z_0 = 0.0$:	$P(0.0 \leq Z \leq \infty) = 0.5000$
$Z_0 = 0.5$:	$P(0.5 \leq Z \leq \infty) = 0.3085$
$Z_0 = 1.0$:	$P(1.0 \leq Z \leq \infty) = 0.1587$
$Z_0 = 1.5$:	$P(1.5 \leq Z \leq \infty) = 0.0668$
$Z_0 = 2.0$:	$P(2.0 \leq Z \leq \infty) = 0.0228$
$Z_0 = 2.5$:	$P(2.5 \leq Z \leq \infty) = 0.00621$
$Z_0 = 3.0$:	$P(3.0 \leq Z \leq \infty) = 0.00135$
$Z_0 = 4.0$:	$P(4.0 \leq Z \leq \infty) = 0.0000317$

Example 1

Given: The life of a particular electronic component is normally distributed with a mean $\mu = 200$ hours and a standard deviation of a $\sigma_x = 20$ hours.

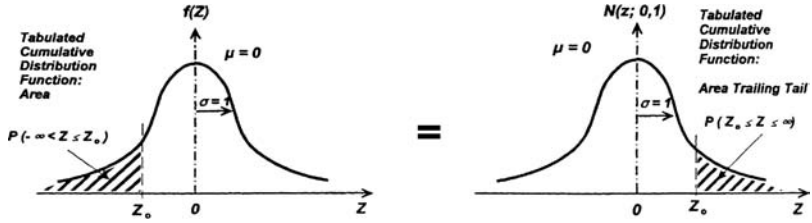


FIGURE 16.22 Electronic components in a symmetrical format of the distribution.

Find: The expected percentage of components requiring replacement at or before 150 hours.

Standard random variable:

$$Z = (X - \mu)/\sigma = \frac{150 - 200}{20} = -2.50$$

That is to say, the specified replacement time (150 hr) is 2.5 σ below the mean.

Cumulative probability distribution:

$$P(-\infty \le X \le 150) = p\,(-\infty \le Z \le -2.5) = 0.00621$$

Hence, only 0.621% of components are expected to be replaced at or before 150 hours.

Because SND is symmetric, you can use either of these tabulations:

$$P(-\infty < Z \le -2.5) = 0.00621 = P(2.5 \le Z < \infty) = 0.00621$$

This may be represented in a pictorial form as in Figure 16.22.

Example 2

Given: A particular engine component is being manufactured having a tension X that is found to be normal distribution with a mean μ = 60 N and a standard deviation of σ_x = 10 N. The manufacturing specifications for the tension of this component are 45 to 70 N.

Find: The percentage of manufactured parts expected to be within specification.

Standard random variable limits:

$$Z_1 = \frac{45 - 60}{10} = -1.50 \quad \text{and} \quad Z_2 = \frac{70 - 60}{10} = 1.00$$

That is to say, the specifications are within 1.5 σ and + 1.0 σ of the manufactured mean. This can be shown in Figure 16.23.

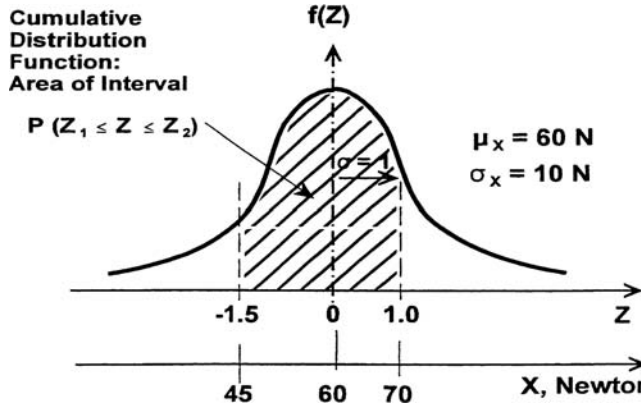


FIGURE 16.23 Area of interval — cumulative distribution function.

Determine net probability:

$$P(45 \le X \le 70) = P(-1.5 \le Z \le 1.0)$$

Cumulate areas consistent with tabulation given. Assume given trailing tail tabulation:

$$\begin{aligned} P(-1.5 \le Z \le 1.0) &= P(-1.5 \le Z \le \infty) - P(1.0 \le Z \le \infty) \\ &= [1 - P(1.5 \le Z \le \infty)] - P(1.0 \le Z \le \infty) \\ &= [1 - 0.0668] - 0.1587 = 0.7745 \end{aligned}$$

Hence, 77.45% of the currently manufactured components are expected to "meet spec." This is shown in Figure 16.24.

So, the reader by now has figured out that at any point in the curve we can indeed find the area provided we know the mean and the standard deviation.

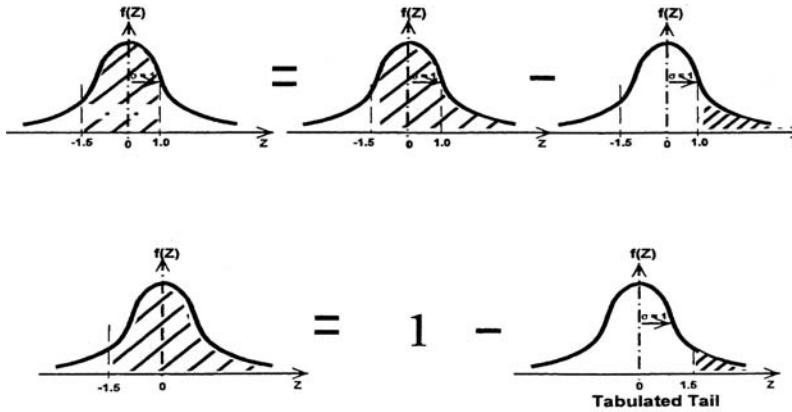


FIGURE 16.24 The graphical progression in figuring out the components of "meeting specifications."

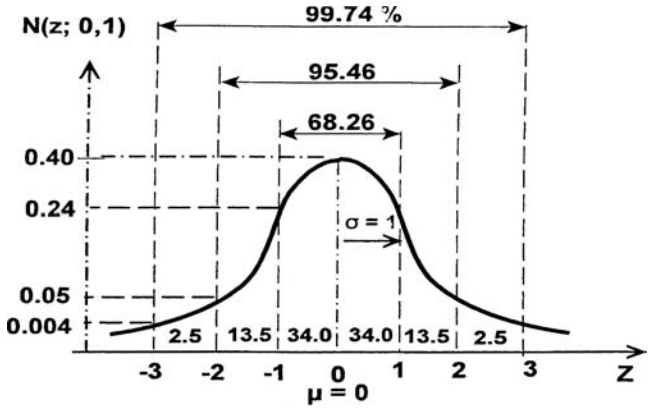


FIGURE 16.25 Percent area under the SND curve.

Figure 16.25 shows the area under the curve in percentage for the traditional mean equals zero and standard deviation of one for the three deviations about the mean.

NORMAL APPROXIMATION OF BINOMIAL

Binomial probability distribution:

$$B(X = x; n, p) = C(x; n)p^x q^{(n-x)} = \frac{n!}{x!(n-x)!} p^x q^{(n-x)}$$

with mean: $\mu = np$ and variance: $\sigma^2 = npq$.

Assumptions:

1. Neither of the binomial probabilities p or q = 1-p is near zero.
2. Sample size n is "sufficiently" large (how large depends upon p).
3. The binomial distribution appears nearly symmetric about the mean in Figure 16.26.

Issues: Binomial lower limit is zero; normal is $-\infty$. Binomial is discrete of cell width, W_c; normal is continuous random variable.

Accuracy improves if:

1. Sample size n increases, or
2. Probability approaches p = 1/2

Certain n and p combinations provide a good approximation for modeling of a binomial population by a normal distribution. Confining our approximation to the interval: $\mu \pm 3\sigma$ requires only that the binomial products $np > 5$ *and* $nq > 5$. These

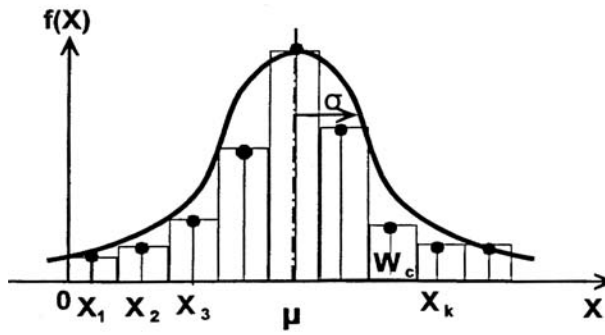


FIGURE 16.26 A typical binomial distribution.

products yield binomial distributions which are nearly symmetric about mean $\mu = np$; and have a variance $\sigma^2 = npq$.

It is this symmetry about the mean that allows us to approximate the binomial distribution as a normal distribution and thereby use the tabulated values to determine its probabilities. Standardized random variables for the binomial distribution are:

$$z = \frac{X - \mu}{\sigma} \approx \frac{X - np}{\sqrt{npq}}$$

After defining the binomial distribution problem, the probability being sought can be determined from the table of standardized normal distribution in the form: $p\ (Z_1 \le Z \le Z_2)$. As a rule of thumb, the following minimum sample sizes may be used for a normal distribution model to approximate a binomial population.

$p = 0.5$	$n \ge 30$
$p = 0.4$ or 0.6	$n \ge 50$
$p = 0.3$ or 0.7	$n \ge 80$
$p = 0.2$ or 0.8	$n \ge 200$
$p = 0.1$ or 0.9	$n \ge 600$
$p = 0.05$ or 0.95	$n \ge 1400$

EXAMPLE 1

Given: A supplier produces a special order of 280 electronic instruments, but past history has shown that 10% of this company's products are defective.

Find: If the number of defects is assumed to be binomially distributed, compute the chance that number of defects does not exceed 23 instruments. (The value 23 is a threshold.)

Statistics:

$$p = 0.10,\ q = 0.90,\ \mu = np = 28 > 5,\ nq = 252 >> 5,\ \sigma = (npq)^{1/2} = 5.0$$

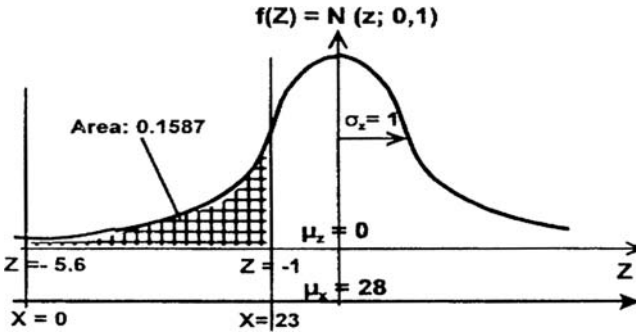


FIGURE 16.27 Normal distribution approximation.

Direct application of binomial distribution

$$P(X \le 23) = \sum_{x=0}^{23} \frac{280!}{X!(280-X)!} p^x q^{280-x}$$

While this is the true calculation, it is very time-consuming, and tables do not exist to condense the labor involved. As a result we use the approximation approach by applying the normal distribution approximation as follows. (In a pictorial format it is shown in Figure 16.27).

$$\begin{aligned} P(0 \le X \le 23) &= P(Z_1 \le Z \le Z_2) \\ &= P\left(\frac{0-28}{5.0} \le Z \le \frac{23-28}{5.0}\right) \\ &= P(-5.6 \le Z \le -1.0) \\ &\simeq P(-\infty \le Z \le -1.0) \\ &= 0.1587 \end{aligned}$$

Therefore, there is a 15.9% chance that *no more than* 23 instruments will not "meet spec."

Some general comments on continuity correction:

1. The normal distribution is based on a continuous RV while the binomial distribution is based on a discrete countable integer RV grouped into cells of a specified width, C_w.
2. The range interval for the continuous normal distribution is from $-\infty$ to $+\infty$, whereas the binomial distribution is on countable integers (items) ranging from 0 to n.
3. A continuity correction is applied to the discrete cells to improve the accuracy of the approximation.

4. The correction consists of:
   a. subtract 1/2-cell width from the lower limit and
   b. add 1/2-cell width to the upper limit
5. These limits in standardized form are:

$$Z_1 = \frac{\left(X_1 - \frac{1}{2}C_w\right) - \mu}{\sigma} \quad \text{and} \quad Z_2 = \frac{\left[\left(X_2 - \frac{1}{2}C_w\right) - \mu\right]}{\sigma}$$

When using the continuity correction, always remember that the error for the normal distribution to approximate the binomial distribution never exceeds:

$$\varepsilon < \frac{0.140}{\sqrt{npq}} = \frac{0.140}{\sigma}$$

EXAMPLE 2

Given: A manufacturer produces a special order of 280 electronic instruments, but past history has shown that 10% of the instruments that it produces are defective.

Find: if the number of defects is assumed to be binomially distributed, compute the chance that exactly 28 defective instruments are manufactured.

Statistics: $p = 0.10$, $q = 0.90$, $\mu = np = 28 > 5$, $n\,q = 252 >> 5$, $\sigma = (npq)^{1/2} = 5.0$, cell width = 1 product, 1/2-cell = 0.5.

Solution:

Direct application of binomial distribution:

$$B(28;280,0.1) = \frac{n!}{X!(n-x)!}p^x q^{n-x} = \frac{280}{28!(280-28)!}(0.1)^{28}(0.9)^{280-28}$$

Normal distribution approximation:

$$\begin{aligned} P((28-0.5) \le X \le (28+0.5)) &= P(Z_1 \le Z \le Z_2) \\ &= P\left(\frac{(27.5)-28}{5.0} \le Z \le \frac{(28.5)-28}{5.0}\right) = P(-0.1 \le Z \le 0.1) \\ &= P(-\infty \le Z \le 0.1) - P(-\infty \le Z \le -0.1) \\ &= 0.5398 - 0.4602 = 0.0796 \end{aligned}$$

Therefore, there is a 7.96% possibility of producing exactly 28 defects.

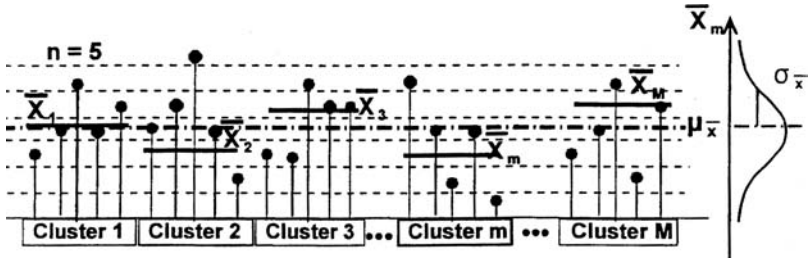


FIGURE 16.28 Mean of the means.

CENTRAL LIMIT THEOREM (CLT) — MEAN OF MEANS IS NORMAL (FIGURE 16.28)

- Statistical inference based on normal distribution.
- Estimation techniques based on normal distribution.
- Real data distribution may not be normal.
- Work with mean of sample clusters, not individual values X_i.
- CLT uses normal distribution to infer population parameter: Mean μ and Variance σ^2

Mathematically the mean of means may be represented by

$$\mu_{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \ldots \bar{x}_M}{M} = \mu$$

Whereas the variance of the means is represented as:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

where n = number of individual samples in a subject or cluster. If there are clusters, the M = total number of clusters, nM = N = total number of individual samples.

Comments on the SND

$$Z \equiv \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

For a cluster of n samples, we can use SND to determine:

1. The probabilities of the sample average, $\overline{X}_m$ or,
2. The required number of samples, n, in a cluster such that is observed mean X_m is within a specified range around the true population mean μ.
   - The cluster size n can be quite small, and the histogram of cluster mean values, X_m, will rapidly converge to a normal distribution regardless of the underlying population.
   - The Central Limit Theorem applies to any population distribution, including the discrete and continuous distributions as well as bimodal distributions.
   - When discrete sampling is involved, the distribution of averages (i.e., the mean of clusters) must be used.
   - The variance of the means is a measure of the spread of clusters means about the true mean.
   - Variance gets smaller as n increases; the smaller the number of samples in a cluster the larger the variance of the means.

NORMALIZED TRANSFORMS

1. Engineering experiments often produce responses that may not appear normal, however, many of these responses can often be transformed into normally distributed representations.
2. The normal distribution is emphasized because many techniques in statistics are based on the assumption of a normal distribution. Two very typical transformations are Y = In X and Y = $\sqrt{X}$. However, there are many more.
3. It is important to recognize that many seemingly nonnormal distributions can be mathematically transformed into what appears to be a normal distribution.
4. A useful normalization is based on the assumption of a large number of basic measurements or observations as well as the use of the Central Limit Theorem.

DISCRETE PROBABILITY DISTRIBUTIONS

Binomial Distribution (Bernoulli)

The binomial and Poisson distributions are the most common distributions with many applications. Key characteristics of the Binomial and Poisson distributions are:

Binomial distribution
- Frequently used in engineering
- Probability of success p and failure q
- Combinations of p's and q's

Poisson Distribution
- Rare successes, p very small
- Large sample size
- Limit of binomial

Population Parameters: characteristics of the population
- Population size: N
- Probability of success: p
- Probability of failure: $q = 1 - p$
- Mean: μ
- Variance: σ^2

Sample Statistics: characteristics of the sample
- Random variable: X_i
- Sample size: n
- Mean: $\overline{X}$
- Variance: s^2

Samples taken without replacement are generally not independent since p is not constant. If sample size $n < 0.05$ N population size, we can consider p unchanged and "independent."

Binomial distribution — Probability when EXACTLY x out of n events occur:

$$B(X_i = x; n, p) \equiv C(x; n)p^x q^{(n-x)}$$

$$= \frac{n!}{x!(n-x)!} p^x q^{(n-x)}; \quad x = 0, 1, \ldots, n$$

The assumptions are:

1. Experiment of n independent events.
2. Probability of a "success" is p.
3. Probability p is constant for all events.
4. Probability of "failure" is $q = 1 - p$.
5. Parameters n and p are specified.
6. Random variable × number of successes.
7. Random variable is a discrete integer $0 \le x \le n$.
8. Order of success not important; combination.

Mean:

$$\mu = E[x] = \sum_{x=0}^{n} xB(x; n, p)$$

$$= np$$

Variance:

$$\sigma^2 = E[(x-\mu)^2] = \sum_{k=0}^{n} (X-\mu)^2 B(x;n,p)$$

$$= npq$$

EXAMPLE 1

Six tosses of a coin.

Experiment is six tosses of a coin, n = 6.

Probability of a head in one toss, p = 1/2.

Probability of a tail in one toss, q = (1 – p) = 1/2.

Find: probability of getting *exactly* 2 heads in 6 tosses.

$$B(x_i = 2;6,1/2) \equiv C(2;6)(1/2)^2(1/2)^{(6-2)}$$

$$= \frac{6!}{2!4!}(1/2)^2(1/2)^4$$

$$= \frac{6\cdot 5}{2\cdot 1}(1/2)^6 = 15(1/2)^6 = \frac{15}{64}$$

Parameters: n = 6 coin tosses, a head with p = 1/2

Random variable: x = 0, 1, 2, 3, 4, 5, 6

r.v.x.	C(x ; n)	$p^x\ q^{n-x}$	B(x;n,p)
0	(6!)/(0! 6!) = 1	$(1/2)^0(1/2)^6 = 1/64$	1/64
1	(6!)/(1! 5!) = 6	$(1/2)^1(1/2)^5 = 1/64$	6/64
2	(6!)/(2! 4!) = 15	$(1/2)^2(1/2)^4 = 1/64$	15/64
3	(6!)/(3! 3!) = 20	$(1/2)^3\ (1/2)^3 = 1/64$	20/64
4	(6!)/(4! 2!) = 15	$(1/2)^4(1/2)^2 = 1/64$	15/64
5	(6!)/(5! 1!) = 6	$(1/2)^5(1/2)^1 = 1/64$	6/64
6	(6!)/(6! 0!) = 1	$(1/2)^6(1/2)^0 = 1/64$	1/64
			Sum = 64/64

If we were to graph this data, we would obtain the graph shown in Figure 16.29.

Parameters: n = 6; p = 1/2; q = 1/2

Mean: $\mu = np = 6 \cdot 1/2 = 3$

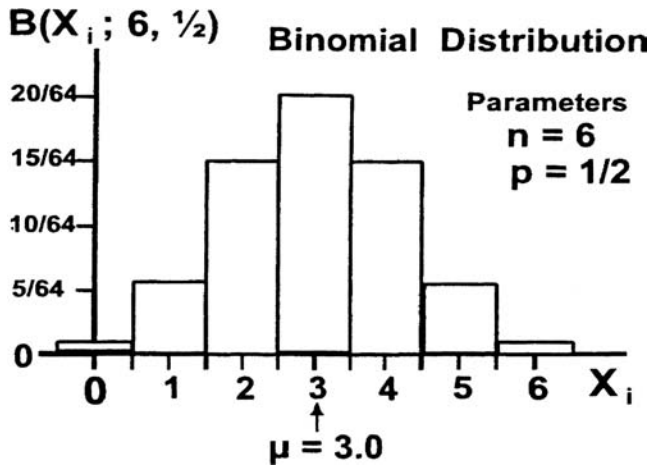


FIGURE 16.29 Binomial distribution histogram: six tosses of a coin.

Variance: $\sigma^2 = npq = 6 \cdot 1/2 \cdot 1/2 = 3/2 = 1.5$

Note that because p = 1/2, (1) the mean is equal to the mid-range, and (2) probability density is symmmetric about the mean.

EXAMPLE 2

Square rod with four sides.

Sides are denoted 1, 2, 3, 4 respectively.

Experiment is six tosses of four-sided rod, n = 6.

Random variable is tossing the number 3.

Probability of a 3 for a single toss, p = 1/4.

Probability of "no 3" in a single toss, q = (1 – p) = 3/4.

Find: Probability of exactly two 3s in 6 tosses; x = 2.

$$B(X_i = 2;6,1/4) \equiv C(2;6)(1/4)^2(3/4)^{(6-2)}$$

$$= \frac{6!}{2!4!}(1/4)^2(3/4)^4$$

$$= \frac{6 \cdot 5}{2 \cdot 1}(1/4)^6(3)^4 = \frac{15 \cdot 3^4}{4^6} = 0.2966$$

Parameters: n = 6 tosses; success a "3," has p = 1/4.

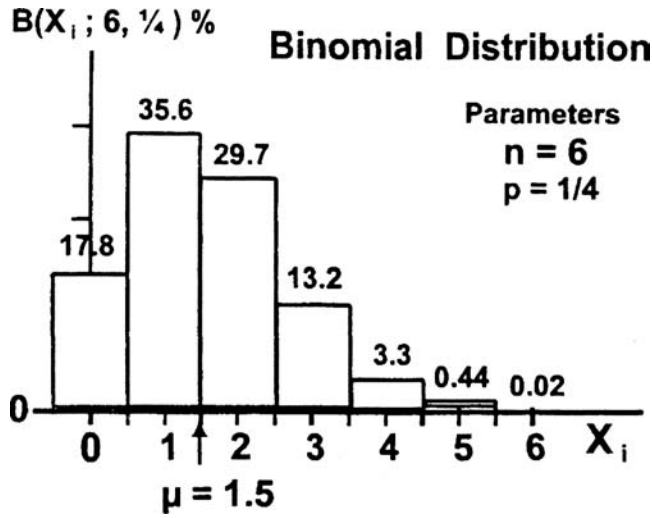


FIGURE 16.30 Histogram in percent of B(x;n,p).

Random variable: x = 0, 1, 2, 3, 4, 5, 6

r.v.x.	C(x;n)	$p^x q^{n-x}$	B(x;n,p)	%
0	(6!)/(0! 6!) = 1	$(1/4)^0\ (3/4)^6 = 3^6/4^6$	$729/4^6$	17.8
1	(6!)/(1! 5!) = 6	$(1/4)^1(3/4)^5 = 3^5/4^6$	$1458/4^6$	35.6
2	(6!)/(2! 4!) = 15	$(1/4)^2(3/4)^4 = 3^4/4^6$	$1215/4^6$	29.7
3	(6!)/(3! 3!) = 20	$(1/4)^3(3/4)^3 = 3^3/4^6$	$540/4^6$	13.2
4	(6!)/(4! 2!) = 15	$(1/4)^4(3/4)^2 = 3^2/4^6$	$135/4^6$	3.30
5	(6!)/(5! 1!) = 6	$(1/4)^5(3/4)^1 = 3^1/4^6$	$18/4^6$	0.44
6	(6!)/(6! 0!) = 1	$(1/4)^6(3/4)^0 = 1/4^6$	$1/4^6$	0.02
Where $4^6 = 4096$			SUM = $4096/4^6$	100

The histogram for this data is shown in percent of B(x;n,p) in Figure 16.30.

Parameters: n = 6, p = 1/4, q = 3/4

Mean: $\mu = np = 6 \cdot 1/4 = 3/2 = 1.5$

Variance: $\sigma^2 = npq = 6 \cdot 1/4 \cdot 3/4 = 18/16 = 1.125$

Observations for p = 1/4: (1) the mean is skewed to the left of the mid-range, and (2) the probability density is nonsymmetric about the mean.

Example 3

Square rod with four sides.

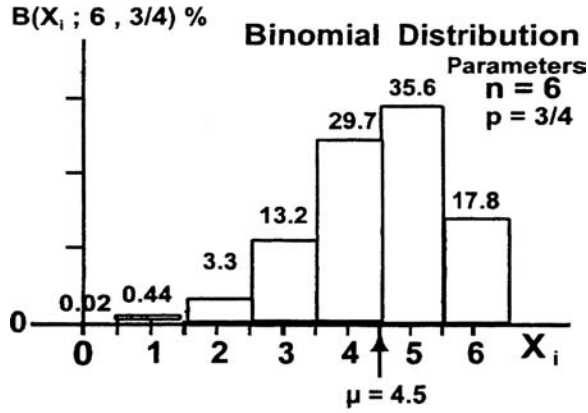


FIGURE 16.31 Binomial distribution for square rod.

Sides are denoted 1, 2, 3, 4, respectively.

Experiment is six tosses of four-sided rod, n = 6.

Random variable is tossing any number except 3.

Probability of "no 3" in a single toss, p = 3/4.

Probability of a 3 for a single toss, q = (1 – p) = 1/4.

The data are shown graphically in Figure 16.31.

Parameters: n = 6, p = 3/4, q = 1/4

Mean: $\mu = np = 6 \cdot 3/4 = 18/4 = 4.5$

Variance: $\sigma^2 = n\ p\ q = 6 \cdot 3/4 \cdot 1/4 = 18/16 = 1.125$

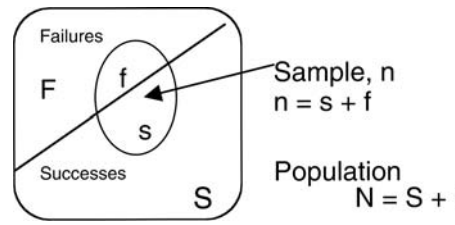
Observations for p = 3/4: (1) the mean is skewed to the right of the mid-range, (2) the probability density is nonsymmetric about the mean.

Hypergeometric Distribution

Overview

Traditional Notation

N = Total population
S = Population defined "success"
F = Population defined "failure"



Proability Density Function (p.d.f.)
This is shown mathematically as:

$$h(x;n,S,N)=\begin{cases}\dfrac{C(x;S)C(n-x;F)}{C(n;N)} & x=0,1,\ldots,n\\ 0 & \text{elsewhere}\end{cases}$$

Note: If total number of successes S < n then x = 0, …, S, then the cumulative distribution frunction (CDF) is shown as:

$$F(x_T)=\sum_{x=0}^{x_T} h(x;n,S,N)$$

General Comments

1. Hypergeometric distribution applies to discrete samples taken from a finite population N without replacement.
2. An integer r is often used in place of the variable x.
3. Three parameters (N, S, and n) specify this distribution.
4. Three alternate parameters (N, p, and n), where (p = S/N), may be used to define the distribution.

Alternate Parameters and Properties

1. Three alternate parameters (N, p, and n), where p = S/N is proportion of success, q = F/N = 1 – p is proportion of failures
2. Hypergeometric distribution (without replacement)

$$h(x;n,p,N)=\frac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}}$$

3. Alternative representation of mean and variance

   Mean: $\mu = np$

   Variance: $\sigma^2 = npq\left(\dfrac{N-n}{N-1}\right)$

Comments

1. As population size N increases (N – n)/(N – 1) goes to 1, and hypergeometric distribution approaches that of binomial.

2. Best unbiased estimator of parameter p from actual data: $p \approx \hat{p} = S / N$; population N not sample n.

Comparison of Hypergeometric and Binomial Distributions

Sample Size

Hypergeometric distribution is based on finite size population N with sample size n taken without replacement. Consequence: Probabilities can vary with sample size n. Binomial and Poisson distributions assume either a finite population N with sample n taken with replacement or a very large population N. Consequence: probabilities are those of population and do not vary with sample size n.

Discrete Two Options

Both hypergeometric and binomial distributions are based on only two kinds of outcomes: pass or fail.

Computations

Hypergeometric probability computations can be quite involved even for small populations. On the other hand, binomial probability or Poisson provide good approximations to hypergeometric and are more easily computed. For a fixed sample size n, this approximation improves with increasing population size N.

Example 1

A supplier provides N = 25 precision motors; history shows that "on average," this company has 8% defects. An inspection sample of n = 8 motors is tested. The RV S is the successful selection of a defect or bad motor; F is a good motor.

$$B = S = 25(0.08) = 2,\ G = F = 25(0.92) = 23,\ n = 8$$

The probability of selecting no defective motors (i.e., x = 0) is:

$$P(X = 0) = \frac{C(0;2)C(8;23)}{C(8;25)}$$

$$= \frac{[2!/0!2!][23!/8!15!]}{[25!/8!17!]} = \frac{17 \cdot 16}{25 \cdot 24} = 0.4533$$

The probability of selecting exactly one defective motor (i.e., x = 1) is:

$$P(X = 1) = \frac{C(1;2)C(7;23)}{C(8;25)}$$

$$= \frac{[2!/1!1!][23!/7!16!]}{[25!/8!17!]} = \frac{2 \cdot 8 \cdot 17}{25 \cdot 24} = 0.4533$$

The probability of selecting exactly two defective motors (i.e., x = 2) is:

$$P(X = 2) = \frac{C(2;2)C(6;23)}{C(8;25)}$$

$$= \frac{[2!/2!0!][23!/6!17!]}{[25!/8!17!]} = \frac{8 \cdot 7}{25 \cdot 24} = 0.0933$$

The probability of selecting fewer than two defective motors from a sample of eight is:

$$P(X < 2) = P(X = 0) + P(X = 1)$$
$$= 0.4533 + 0.4533 = 0.9066$$

That is, there is a 90% certainty of observing fewer than two bad motors in a total sample size of eight.

Here is how to determine the mean and standard deviation of the number of defective motors in a sample size of n = 8 and of n = 12:

For n = 8:

$$\mu = np = 8(0.08) = 0.64 \text{ motors}$$

$$\sigma^2 = npq\left(\frac{N-n}{N-1}\right) = 8(0.08)(0.92)\left(\frac{25-8}{25-1}\right) = 0.4171$$

$$\sigma = 0.65 \text{ motors}$$

For n = 12:

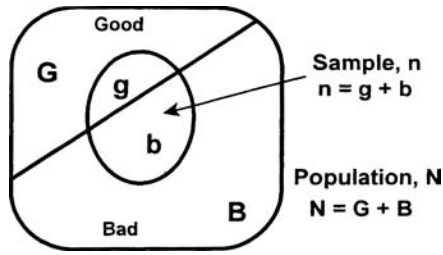
$$\mu = np = 12(0.08) = 0.96 \text{ motors}$$

$$\sigma^2 = npq\left(\frac{N-n}{N-1}\right) = 12(0.08)(0.92)\left(\frac{25-12}{25-1}\right) = 0.4784$$

$$\sigma = 0.69 \text{ motors}$$

Hypergeometric Distribution Applications

Often, the meaning of the terms "success" and "failure" depends upon the context in which they are used. For example, selecting a "bad" motor could be considered a "success" for their removal. When there is an issue of success or failure or "bad" or "good," the hypergeometric distribution is applicable. For example, in the following sample space we have the good and bad motors for a given supplier segregated as shown:



A supplier delivers a total of N motors. This population of N motors is comprised of only two classes:

1. G motors that will be "good" or pass the specification.
2. B (= N – G): motors that will be "bad" or fail specification.

Quality control inspects only a small sample of n motors

g motors are good
b (= n – g) motors are bad

Probability Considerations

1. The inspection sample of n motors is taken without replacement from the finite population of N motors.
2. Without replacement, the probabilities of successes and failures for each of inspection sample of n motors will not be constant but will depend upon what motors are selected.
3. An inspection sample of n motors can have b bad motors and g good motors, where n = g + b.
4. Using combination theory, the total number of unordered ways of selecting g good motors from a population containing a total of G good motors is:

$$C(g;G) = \frac{G!}{g!(G-g)!}$$

5. Also, the number of unordered ways of selecting b bad motors from a population containing a total of B bad motors is the combination:

$$C(b;B) = B!/b!(B-b)!$$

6. The total number of ways to get both b bad motors and g good motors is the product:

$$C(g;G)C(b;B) = \frac{G!}{g!(G-g)!}\frac{B!}{b!(B-b)!}$$

7. The total number of ways of selecting n motors from a population N (without replacement and unordered) is:

$$C(n;N) = \frac{N!}{n!(N-n)!}$$

8. Quality inspectors assume that it is "equally likely" to select any sample of n motors. That is, a sample of n motors containing b_1 bad and $g_1 = (n - b_1)$ good motors is as likely to be selected as a sample of n-motors containing b_2 bad and $g_2 = (n - b_2)$ good motors.
9. The probability of any combination of samples containing n motors is given by:

$$P(\text{each sample}) = \frac{1}{\text{total number of sample combinations}}$$

$$= \frac{1}{C(n;N)}$$

10. The probability of selecting a sample of n-motors containing exactly g good motors and b bad motors is given by:

$$P(g - \text{good},\ \ b - \text{bad}) = \frac{\text{number of n } - \text{ samples with g and b}}{\text{total number of sample combinations}}$$

$$= \frac{C(g;G)C(b;B)}{C(n;N)}$$

[Recall: If independent $P(A \cdot B) = P(A|B)\ P(B) = P(A)\ P(B)$]

Random Variable

Generally a random variable X is used to define or describe a specific form of outcome. In our case we consider the random variable to be the "success" of selecting a number of bad motors from an inspection sample of n motors. (Note: "success" = bad.)

Quality inspectors set a threshold for the value of this random variable, say $X = x_b$, that has be established to assure the delivered lot of N motors will be accepted. (This does not mean that all N motors "meet spec," only that some acceptable percentage do.)

$x_b = b$ represents some (integer value) lower limit for acceptable. The number of "failures" or good motors in sample is $g = n - x_b$. The probability of exactly x motors passing in the sample of n motors is:

$$P(X = x_b) = \frac{C(x_b;B)C(n - x_b;G)}{C(n;N)}$$

Example 2

A supplier provides N = 10 precision motor;, history shows that "on average," this company's products have 10% defects. If an inspection sample of n = 1 motor is tested,

what is the probability that exactly one defective motor is selected (i.e., $x_b = 1$)? The RV X is the "successful" selection of a bad motor; G is a good motor.

$$x = B = 10(0.1) = 1, \qquad G = 10(0.9) = 9, \qquad n = 1, \qquad x_b = 1$$

$$P(X = x_b = 1) = \frac{C(1;1)C(0;9)}{C(1;10)}$$

$$= \frac{[1!/1!0!][9!/0!9!]}{[10!/1!9!]} = \frac{1}{10} = 0.1000$$

Example 3

If an inspection sample of n = 2 motors is tested, what is the probability that exactly one defective motor is selected (i.e., $x_b = 1$)?

$$X = B = 1, \qquad G = 9, \qquad N = 10, \qquad n = 2, \qquad x_b = 1$$

$$P(X = x_b = 1) = \frac{C(1;1)C(1;9)}{C(2;10)}$$

$$= \frac{[1!/1!O!][9!/1!8!]}{[10!/2!8!]} = \frac{2}{10} = 0.2000$$

This implies that with two selections the inspector has a 20% chance of detecting the bad motor.

As the number of samples increases, n = 3, 4, 5, …, N, the probability of selecting the one defective motor increases as 0.1n. So for five samples the probability is 50% that the inspector will have selected one bad motor.

Example 4

In the previous examples we designated a success as the selection of exactly one bad or defective motor. It may appear counterintuitive to identify a "success " with a "defect." However, examine what would happen if we defined a "success" X as selecting one "good" motor from a sample size of three.

$$X = G = 9, \qquad B = 1, \qquad N = 10, \qquad n = 3, \qquad x_g = 1$$

$$P(X = x_g = 1) = \frac{C(1;9)C(2;1)}{C(3;10)}$$

However, C(2; 1) is undefined since we cannot have (–1)!. The reason, of course, is that we only have one bad motor, and by only asking for one "success" or good motor to be selected in a sample of three means the other two selections must be bad motors. But because there is only a total of one bad motor, it is impossible to satisfy this probability. We exceeded our expectations since we only have one "bad" motor.

POISSON DISTRIBUTION: LIMIT OF BINOMIAL DISTRIBUTION FOR RARE OCCURRENCE

Rare successes: Probability p extremely small but sample size n extremely large, such that binomial mean = np = a is finite. Since p is so small, we can expect that x success will also be small compared to sample size n. The Poisson distribution is also used to compute the probability of the number of "Poisson events" during a given time interval or within a specified region of space.

We already have defined the binomial distribution as:

$$B(X_i = x; n, p) \equiv C(x; n)p^x q^{(n-x)}$$

$$= \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)}$$

For the Poisson distribution, however, we:

1. Reduce the permutation factor, since x << n.

$$\frac{n!}{(n-x)!} = n(n-1)\ldots(n-(x-1)$$

$$\approx n(n)\ldots(n) = n^x$$

2. Combine factors of like powers.

$$B(X_i = x; n, p) \equiv \frac{n^x}{x!} p^x (1-p)^{(n-x)}$$

$$= \frac{(np)^x}{x!} \frac{(1-p)^n}{(1-p)^x}$$

3. Note that because p « 1 and x is not large, we can approximate $(1-p)^x = 1$.
4. Having mean np = a we can rewrite n = a/p so

$$B(X_i = x; n, p) \equiv \frac{a^x}{x!}(1-p)^{a/p}$$

$$= \frac{(a)^x}{X!}[(1-p)^{1/p}]^a$$

5. Limit as $p \to 0$ yields exponential

$$\lim_{p\to 0}(1-p)^{1/p} = \frac{1}{e} = e^{-1}$$

6. The result is the Poisson distribution (one parameter: a).

$$P_0(X_i = x;a) \equiv \frac{a^x}{x!}e^{-1}$$

   a. In reliability, the parameter $a = \delta t$.
   b. δ is the mean number of events per unit time.

Comparison of Binomial and Poisson

The probability density function (pdf) for the binomial distribution is:

$$B\ (X_i = x;n,\ p) = [n!/X!(n - x)!]\ p^x q^{(n-x)}$$

Mean: $\mu = np = a$
Variance: $\sigma^2 = npq = aq$
Two Parameters: n and p
or
a(= np) and q = (1– p)
On the other hand, the probability function for the Poison distribution is:

$$P_o(X_i = x;a) \equiv \frac{a^x}{x!}e^{-a}$$

where a = n p and q = (1 – p) approx. 1.
Mean: $\mu = a$
Variance: $\sigma^2 = a$
One parameter: a(= np)

Example 5

A compressor manufacturer has a record of 50 defects per 1000 produced, which corresponds to a defect percentage of 5%. Since the probability of successfully observing a defect is small (p = 0.05), then we can assume a Poisson distribution.

Determine probability of observing defects in batch of n = 10:

Poisson parameter: a = np = 10(0.05) = 0.5

Probability of exactly x = 0 failures:

$$P_o(X = 0;0.5) \equiv \frac{(0.5)^0}{0!}e^{-0.5} = \frac{1}{1}(0.6065) = 0.6065$$

Probability of exactly x = 1 failure:

$$P_0(X = 1; 0.5) \equiv \frac{(0.5)^1}{1!} e^{-0.5} = \frac{0.5}{1}(0.6065) = 0.3033$$

Probability of exactly x = 2 failures:

$$P_0(X = 2; 0.5) \equiv \frac{(0.5)^2}{2!} e^{-0.5} = \frac{0.25}{2!}(0.6065) = 0.0758$$

Probability of exactly x = 3 failures:

$$P_0(X = 3; 0.5) \equiv \frac{(0.5)^3}{3!} e^{-0.5} = \frac{0.125}{3 \cdot 2}(0.6065) = 0.00158$$

Figure 16.32 shows the Poisson distribution for each of the failures.

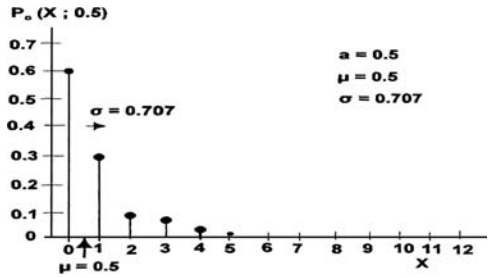
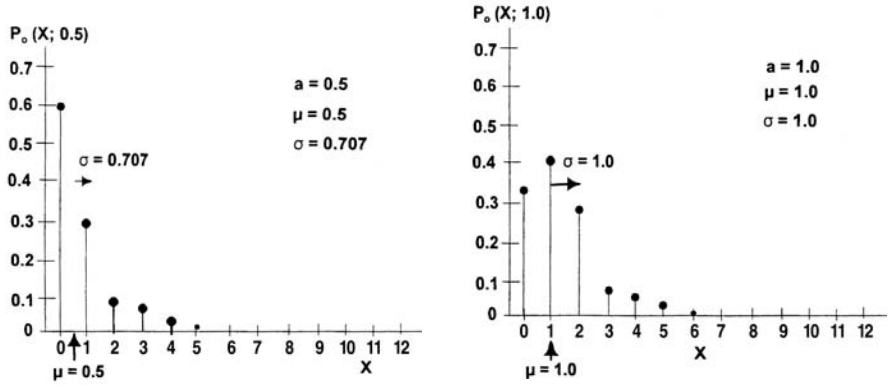
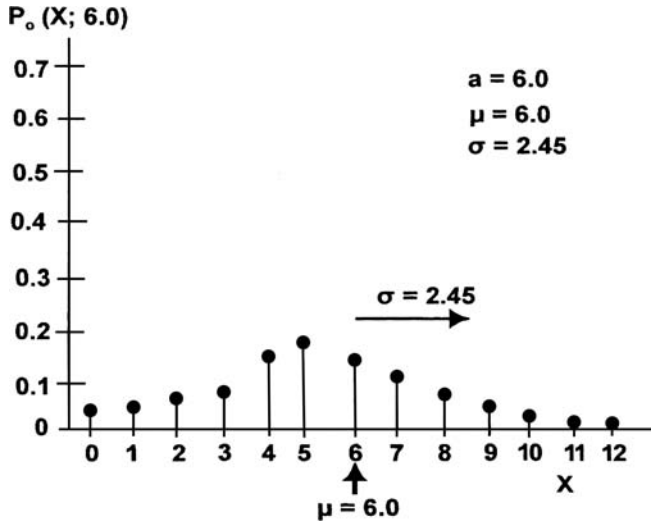


FIGURE 16.32 Poisson distribution for the four failures.



Special comments:

1. The mean is equal to the parameter a; as a result, the smaller the "a's," the more skewed the probability density function (pdf) is to the left.



2. The standard deviation is equal to the square root of "a"; hence, increasing the value of "a" also increases the standard deviation.
3. The larger the value of "a," the more the Poisson distribution approaches the normal distribution.

SELECTED BIBLIOGRAPHY

Becker, W.E., *Business and Economics Statistics*, Addison-Wesley Publishing Co., Reading, MA, 1987.

Cox, D.R. and Snell, E.J., *Applied Statistics*, Chapman & Hall, New York, 1981.

Daniel, W.W. and Terrell, J.C., *Business Statistics for Management and Economics,* Houghton Mifflin Co., Boston, 1989.

Deming, W.E., *Some Theory of Sampling*, Dover Publications, New York, 1966.

Freund, J.E. and Williams, F.J., *Elementary Business Statistics: The Modern Approach*, 2nd ed., Prentice Hall, Englewood Cliffs, NJ, 1972.

Fuller, W.A., *Measurement Error Models*, John Wiley & Sons, New York, 1987.

Gibbons, G.D., *Nonparametric Statistical Inference*, 2nd ed., Marcel Dekker, New York, 1985.

Hays, W., *Statistics*, 3rd ed., Holt, Rinehart & Winston, New York, 1981.

Huck, S.W., Cormier, W.H., and Bounds, W.G., Jr., *Reading Statistics and Research*, Harper & Row, New York, 1974.

Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H., and Lee, T.C., *The Theory and Practice of Econometrics*, 2nd ed., John Wiley & Sons, New York, 1985.

Lehmann, E.L., *Testing Statistical Hypothesis*, John Wiley & Sons, New York, 1986.

Mansfield, E., *Statistics for Business and Economics: Methods and Applications*, 2nd ed., W.W. Norton & Company, New York, 1983.

Stuart, A., *Statistics and Probability,* notes from a Ford Motor Company training seminar, Lemont, PA, 200.

Part III

Appendices

Appendix A — Matrix Algebra: An Introduction

Matrix algebra is one of the most useful and powerful branches of mathematics for conceptualizing and analyzing engineering, psychological, sociological, and educational research data. As research becomes more and more multivariate, the need for a compact method of expressing data becomes greater. Certain problems require that sets of equations and subscripted variables be written. In many cases, the use of matrix algebra simplifies and, when familiar, clarifies the mathematics and statistics. In addition, matrix algebra notation and thinking fit in nicely with the conceptualization of computer programming and use.

This appendix provides a brief introduction to matrix algebra. The emphasis is on those aspects that are related to subject matter covered in this complete series. Thus many matrix algebra techniques, important and useful in other contexts, are omitted. In addition, certain important derivations and proofs are neglected. Although the material presented here should suffice to enable you to follow the applications of matrix algebra in this series especially Volumes III and V — it is strongly suggested that you expand your knowledge of this topic by studying one or more of the following texts: Dorf (1969), Green (1976), Hohn (1964), Horst (1963), Searle (1966), and Strang (1980).

BASIC DEFINITIONS

A *matrix* is an n-by-k rectangle of numbers or symbols that stand for numbers. The order of the matrix is n by k. It is customary to designate the rows first and the columns second. That is, n is the number of rows of the matrix and k the number of columns. A 2-by-3 matrix called **A** might be

$$\mathbf{A} = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 4 & 7 & 5 \\ 6 & 6 & 3 \end{bmatrix} \end{array}$$

Elements of a matrix are identified by reference to the row and column that they occupy. Thus, a_{11} refers to the element of the first row and first column of **A**, which in the above example is 4. Similarly, a_{23} is the element of the second row and third column of **A**, which in the above example is 3. In general, then, a_{ij} refers to the element in row i and column j.

The *transpose* of a matrix is obtained simply by exchanging rows and columns. In the present case, the transpose of **A**, written **A'**, is

$$\mathbf{A}' = \begin{bmatrix} 4 & 6 \\ 7 & 6 \\ 5 & 3 \end{bmatrix}$$

If n = k, the matrix is square. A square matrix can be symmetric or asymmetric. A *symmetric* matrix has the same elements above the principal diagonal as below the diagonal except that they are transposed. The principal diagonal is the set of elements from the upper left corner to the lower right corner. Symmetric matrices are frequently encountered in multiple regression analysis and in multivariate analysis. The following is an example of a correlation matrix, which is symmetric:

$$\mathbf{R} = \begin{bmatrix} 1.00 & .70 & .30 \\ .70 & 1.00 & .40 \\ .30 & .40 & 1.00 \end{bmatrix}$$

Diagonal elements refer to correlations of variables with themselves, hence the 1's. Each off-diagonal element refers to a correlation between two variables and is identified by row and column numbers. Thus, $r_{12} = r_{21} = .70$; $r_{23} = r_{32} = .40$. A column *vector* is an n-by-1 array of numbers. For example:

$$\mathbf{b} = \begin{bmatrix} 8.0 \\ 1.3 \\ -2.0 \end{bmatrix}$$

A *row vector* is a 1-by-n array of numbers:

$$\mathbf{b}' = [8.0 \quad 1.3 \quad -.20]$$

b' is the *transpose* of **b**. Note that vectors are designated by lowercase boldface letters, and that a prime is used to indicate a row vector.

A *diagonal* matrix is frequently encountered in statistical work. It is simply a matrix in which some values other than zero are in the principal diagonal of the matrix, and all the off-diagonal elements are zeros. Here is a diagonal matrix:

$$\begin{bmatrix} 2.759 & 0 & 0 \\ 0 & 1.643 & 0 \\ 0 & 0 & .879 \end{bmatrix}$$

A particularly important form of a diagonal matrix is an *identity* matrix, **I**, which has ones in the principal diagonal:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

MATRIX OPERATIONS

The power of matrix algebra becomes apparent when we explore the operations that are possible. The major operations are addition, subtraction, multiplication, and inversion. Many statistical operations can be done by knowing the basic rules of matrix algebra. Some matrix operations are now defined and illustrated.

Addition and Subtraction

Two or more vectors can be added or subtracted provided they are of the same dimensionality. That is, they have the same number of elements. The following two vectors are added:

$$\begin{bmatrix} 4 \\ 3 \\ 5 \end{bmatrix} + \begin{bmatrix} 7 \\ 7 \\ 4 \end{bmatrix} = \begin{bmatrix} 11 \\ 10 \\ 9 \end{bmatrix}$$

a **b** **c**

Similarly, matrices of the same dimensionality may be added or subtracted. The following two 3-by-2 matrices are added:

$$\begin{bmatrix} 6 & 4 \\ 5 & 6 \\ 9 & 5 \end{bmatrix} + \begin{bmatrix} 7 & 4 \\ 7 & 4 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 13 & 8 \\ 12 & 10 \\ 10 & 8 \end{bmatrix}$$

A **B** **C**

Now, **B** is subtracted from **A**:

$$\begin{bmatrix} 6 & 4 \\ 5 & 6 \\ 9 & 5 \end{bmatrix} - \begin{bmatrix} 7 & 4 \\ 7 & 4 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -2 & 2 \\ 8 & 2 \end{bmatrix}$$

A **B** **C**

Multiplication

To obtain the product of a row vector by a column vector, corresponding elements of each are multiplied and then added. For example, the multiplication of **a**' by **b**, each consisting of three elements, is:

$$\underset{\mathbf{a}'}{[a_1 \quad a_2 \quad a_3]}\underset{\mathbf{b}}{\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}} = a_1b_1 + a_2b_2 + a_3b_3$$

Note that the product of a row by a column is a single number called a scalar. This is why the product of a row by a column is referred to as the scalar product of vectors. Here is a numerical example:

$$[4 \quad 1 \quad 3]\begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix} = (4)(1) + (1)(2) + (3)(5) = 21$$

Scalar products of vectors are very frequently used in statistical analysis. For example, to obtain the sum of the elements of a column vector it is premultiplied by a unit row vector of the same dimensionality. Thus,

$$\Sigma X\text{: } [1 \quad 1 \quad 1 \quad 1]\begin{bmatrix} 1 \\ 4 \\ 1 \\ 3 \\ 7 \end{bmatrix} = 16$$

The sum of the squares of a column vector is obtained by premultiplying the vector by its transpose:

$$\Sigma X^2\text{: } [1 \quad 4 \quad 1 \quad 3 \quad 7]\begin{bmatrix} 1 \\ 4 \\ 1 \\ 3 \\ 7 \end{bmatrix} = 76$$

Similarly, the sum of the products of X and Y is obtained by multiplying the row of X by the column of Y, or the row of Y by the column of X.

$$\Sigma XY: \begin{bmatrix}1 & 4 & 1 & 3 & 7\end{bmatrix}\begin{bmatrix}3\\-5\\7\\2\\-1\end{bmatrix} = -11$$

Scalar products of vectors are used frequently in dealing with multiple regression, discriminant analysis, multivariate analysis of variance, and canonical analysis.

Instead of multiplying a row vector by a column vector, one may multiply a column vector by a row vector. The two operations are entirely different from each other. It was shown above that the former results in a scalar. The latter operation, on the other hand, results in a matrix. This is why it is referred to as the matrix product of vectors. For example:

$$\begin{bmatrix}3\\-5\\7\\2\\-1\end{bmatrix}\begin{bmatrix}1 & 4 & 1 & 3 & 7\end{bmatrix} = \begin{bmatrix}3 & 12 & 3 & 9 & 21\\-5 & -20 & -5 & -15 & -35\\7 & 28 & 7 & 21 & 49\\2 & 8 & 2 & 6 & 14\\-1 & -4 & -1 & -3 & -7\end{bmatrix}$$

Note that each element of the column is multiplied, in turn, by each element of the row to obtain one element of the matrix. The product of the first element of the column by the row elements becomes the first row of the matrix. Those of the second element of the column by the row become the second row of the matrix, and so forth. Thus, the matrix product of a column vector of k elements is a $k \times k$ matrix.

Matrix multiplication is done by multiplying rows by columns. An example is easier than verbal explanation. Suppose we want to multiply two matrices, **A** and **B**, to produce the product matrix **C**:

$$\begin{bmatrix}3 & 1\\5 & 1\\2 & 4\end{bmatrix} \times \begin{bmatrix}4 & 1 & 4\\5 & 6 & 2\end{bmatrix} = \begin{bmatrix}17 & 9 & 14\\25 & 11 & 22\\28 & 26 & 16\end{bmatrix}$$

A **B** **C**

Following the rule of scalar product of vectors, we multiply and add as follows (follow the arrows):

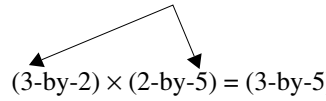
$$(3)(4)+(1)(5)=17 \qquad (3)(1)+(1)(6)=9 \qquad (3)(4)+(1)(2)=14$$

$$(5)(4)+(1)(5)=25 \qquad (5)(1)+(1)(6)=11 \qquad (5)(4)+(1)(2)=22$$

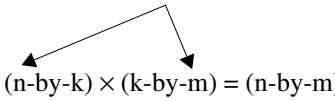
$$(2)(4)+(4)(5)=28 \qquad (2)(1)+(4)(6)=26 \qquad (2)(4)+(4)(2)=16$$

From the foregoing illustration, it may be discerned that in order to multiply two matrices it is necessary that the number of columns of the first matrix be equal to the number of rows of the second matrix. This is referred to as the *conformability* condition. Thus, for example, an n-by-k matrix can be multiplied by a k-by-m matrix because the number of rows of the first (k) is equal to the number of rows of the second (k). In this context, the k's are referred to as the "interior" dimensions; n and m are referred to as the "exterior" dimensions.

Two matrices are conformable when they have the same "interior" dimensions. There are no restrictions on the "exterior" dimensions when two matrices are multiplied. It is useful to note that the "exterior" dimensions of two matrices being multiplied become the dimensions of the product matrix. For example, when a 3-by-2 matrix is multiplied by a 2-by-5 matrix, a 3-by-5 matrix is obtained:



In general,



A special case of matrix multiplication often encountered in statistical work is the multiplication of a matrix by its transpose to obtain a matrix of raw score or deviation Sums of Squares and Cross Products (SSCP). Assume that there are n subjects for whom measures on k variables are available. In other words, assume that the data matrix, **X**, is an n-by-k. To obtain the raw score SSCP, calculate **X'X**. Here is a numerical example:

$$\begin{array}{ccc} & \text{n} & \\ & & \\ \text{K}\begin{bmatrix} 1 & 4 & 137 \\ 2 & 3 & 346 \\ 2 & 5 & 135 \end{bmatrix} & \text{n} & \end{array} \quad \begin{array}{ccc} & \text{k} & \\ & & \\ \begin{bmatrix} 1 & 2 & 2 \\ 4 & 3 & 5 \\ 1 & 3 & 1 \\ 3 & 4 & 3 \\ 7 & 6 & 5 \end{bmatrix} & & \end{array} = \begin{bmatrix} 76 & 71 & 67 \\ 71 & 74 & 64 \\ 67 & 64 & 64 \end{bmatrix}$$

$$\mathbf{X'} \qquad\qquad \mathbf{X} \qquad\qquad \mathbf{X'X}$$

In statistical symbols, **X'X** is

$$\Sigma X_i X_j = \begin{bmatrix} \Sigma X_1^2 & \Sigma X_1 X_2 & \Sigma X_1 X_3 \\ \Sigma X_2 X_1 & \Sigma X_2^2 & \Sigma X_2 X_3 \\ \Sigma X_3 X_1 & \Sigma X_3 X_2 & \Sigma X_2^3 \end{bmatrix}$$

Using similar operations, one may obtain deviation SSCP matrices. Such matrices are used frequently in statistical calculations of advanced methodologies.

A matrix can be multiplied by a scalar: each element of the matrix is multiplied by the scalar. Suppose, for example, we want to calculate the mean of each of the elements of a matrix of sums of scores. Let N = 10. The operation is

$$1/10 \begin{bmatrix} 20 & 48 \\ 30 & 40 \\ 35 & 39 \end{bmatrix} = \begin{bmatrix} 2.0 & 4.8 \\ 3.0 & 4.0 \\ 3.5 & 3.9 \end{bmatrix}$$

Each element of the matrix is multiplied by the scalar 1/10.

A matrix can also be multiplied by a vector. The first example given below is premultiplication by a vector; the second is postmultiplication:

$$[6 \quad 5 \quad 2] \begin{bmatrix} 7 & 3 \\ 7 & 2 \\ 4 & 1 \end{bmatrix} [85 \quad 30]$$

$$\begin{bmatrix} 7 & 7 & 4 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 6 \\ 5 \\ 2 \end{bmatrix} = \begin{bmatrix} 85 \\ 30 \end{bmatrix}$$

Note that in the latter example, (2-by-3) × (3-by-1) becomes (2-by-1). This sort of multiplication of a matrix by a vector is done very frequently in multiple regression.

Thus far, nothing has been said about the operation of division in matrix algebra. In order to show how this is done it is necessary first to discuss some other concepts, to which we now turn.

Determinants

A determinant is a certain numerical value associated with a square matrix. The determinant of a matrix is indicated by vertical lines instead of brackets. For example, the determinant of a matrix **B** is written

$$\text{Det } \mathbf{B} = |\boldsymbol{B}| = \begin{vmatrix} 4 & 2 \\ 1 & 5 \end{vmatrix}$$

B

The calculation of the determinant of a 2×2 matrix is very simple: it is the product of the elements of the principal diagonal minus the product of the remaining two elements. For the above matrix,

$$|\boldsymbol{B}| = \begin{vmatrix} 4 & 2 \\ 1 & 5 \end{vmatrix} = (4)(5) - (1)(2) = 20 - 2 = 18$$

or, symbolically,

$$|\boldsymbol{B}| = \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} = b_{11}b_{22} - b_{12}b_{21}$$

The calculation of determinants for larger matrices is quite tedious and will not be shown here (see references listed at the end of this appendix). In any event, matrix operations are most often done with the aid of a computer. The purpose here is solely to indicate the role played by determinants in some applications of statistical analysis.

Applications of Determinants

To give the flavor of the place and usefulness of determinants in statistical analysis we turn first to two simple correlation examples. Suppose we have two correlation coefficients, r_{y1}, and r_{y2}, calculated between a dependent variable, Y, and two variables, 1 and 2. The correlations are $r_{y1} = .80$ and $r_{y2} = .20$. We set up two matrices that express the two relations, but this is done immediately in the form of determinants, whose numerical values are calculated:

$$\begin{array}{cc} 1 & y \\ \end{array}$$

$$\begin{vmatrix} 1.00 & .80 \\ .80 & 1.00 \end{vmatrix} = (1.00)(1.00) - (.80)(.80) = .36$$

and

$$\begin{array}{cc} 2 & y \\ \end{array}$$

$$\begin{vmatrix} 1.00 & .20 \\ .20 & 1.00 \end{vmatrix} = (1.00)(1.00) - (.20)(.20) = .96$$

The two determinants are .36 and .96. Now, to determine the percentage of variance shared by y and 1 and by y and 2, square the r's:

$$r_{y1}^2 = (.80)^2 = .64$$

$$r_{y2}^2 = (.20)^2 = .04$$

Subtract each of these from 1.00: 1.00 – .64 = .36, and 1.00 – .04 = .96. These are the determinants just calculated. They are $1 - r^2$, or the proportions of the variance not accounted for.

As an extension of the foregoing demonstration, it may be shown how the squared multiple correlation, R^2, can be calculated with determinants:

$$R^2_{y.12\ldots k} = 1 - \frac{|\boldsymbol{R}|}{|\boldsymbol{R}_x|}$$

where $|\boldsymbol{R}|$ is the determinant of the correlation matrix of all the variables, that is, the independent variables as well as the dependent variable; $|\boldsymbol{R}_x|$ is the determinant of the correlation matrix of the independent variables. From the foregoing it can be seen that the ratio of the two determinants indicates the proportion of variance of the dependent variable, Y, not accounted by the independent variables, X's. The ratio of two determinants is also frequently used in multivariate analyses dealing with Wilks' Λ.

Another important use of determinants is related to the concept of linear dependencies, to which we now turn.

Linear Dependence

Linear dependence means that one or more vectors of a matrix, rows or columns, are a linear combination of other vectors of the matrix. The vectors **a'** = [3 1 4] and **b'** = [6 2 8] are dependent since 2**a'** = **b'**. If one vector is a function of another in this manner, the coefficient of correlation between them is 1.00. Dependence in a matrix can be defined by reference to its determinant. If the determinant of the matrix is zero it means that the matrix contains at least one linear dependency. Such a matrix is referred to as being *singular.* For example, calculate the determinant of the following matrix:

$$\begin{vmatrix} 3 & 1 \\ 6 & 2 \end{vmatrix} = (3)(2) - (1)(6) = 0$$

The matrix is singular, that is, it contains a linear dependency. Note that the values of the second row are twice the values of the first row.

A matrix with a determinant other than zero is referred to as being *nonsingular.* The notions of singularity and nonsingularity of matrices play very important roles in statistical analysis, especially in the analysis of multicollinearity. As is shown below, a singular matrix has no inverse.

We turn now to the operation of division in matrix algebra, which is presented in the context of the discussion of matrix inversion.

Matrix Inverse

Recall that the division of one number into another number amounts to multiplying the dividend by the reciprocal of the divisor:

$$\frac{a}{b} = \frac{1}{b}a$$

For example, 12/4 = 12(1/4) = (12)(.25) = 3. Analogously, in matrix algebra, instead of dividing a matrix **A** by another matrix **B** to obtain matrix **C**, we multiply **A** by the *inverse* of **B** to obtain **C**. The inverse of **B** is written $\mathbf{B}^{-1}$. Suppose, in ordinary algebra, we had ab = c, and we wanted to find b. We would write

$$b = \frac{c}{a}$$

In matrix algebra, we write

$$\mathbf{B} = \mathbf{A}^{-1}\mathbf{C}$$

(Note that **C** is premultiplied by $\mathbf{A}^{-1}$ and not postmultiplied. In general, $\mathbf{A}^{-1}\mathbf{C} \neq \mathbf{C}\mathbf{A}^{-1}$.)

The formal definition of the inverse of a square matrix is: Given **A** and **B**, two square matrices, if **AB** = **I**, then **A** is the inverse of **B**.

Generally, the calculation of the inverse of a matrix is very laborious and, therefore, error prone. This is why it is best to use a computer program for such purposes (see below). Fortunately, however, the calculation of the inverse of a 2 × 2 matrix is very simple, and is shown here for three reasons:

1. It affords an illustration of the basic approach to the calculation of the inverse.
2. It affords the opportunity to show the role played by the determinant in the calculation of the inverse.
3. Inverses of 2 × 2 matrices are frequently calculated in some applications of statistical tools and especially in multivariate analysis.

In order to show how the inverse of a 2 × 2 matrix is calculated, it is necessary first to discuss briefly the *adjoint* of such a matrix. This is shown in reference to the following matrix:

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

The adjoint of **A** is:

$$\text{Adj}\,\mathbf{A} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Thus, to obtain the adjoint of a 2 × 2 matrix, interchange the elements of its principal diagonal (a and d in the above example), and change the signs of the other two elements (b and c in the above example). Now the inverse of a matrix **A** is:

$$\mathbf{A}^{-1} = \frac{adjA}{|A|} = \frac{1}{|A|}\text{adj}\,\mathbf{A}$$

where $|\boldsymbol{A}|$ is the determinant of **A**. The inverse of the following matrix, **A**, is now calculated.

$$\mathbf{A} = \begin{bmatrix} 6 & 2 \\ 8 & 4 \end{bmatrix}$$

First, calculate the determinant of **A**:

$$|\boldsymbol{A}| = \begin{vmatrix} 6 & 2 \\ 8 & 4 \end{vmatrix} = (6)(4) - (2)(8) = 8$$

Second, form the adjoint of **A**:

$$\text{adj}\,\mathbf{A} = \begin{bmatrix} 4 & -2 \\ -8 & 6 \end{bmatrix}$$

Third, multiply the adj **A** by the reciprocal of $|\boldsymbol{A}|$ to obtain the inverse of **A**.

$$\mathbf{A}^{-1} = \frac{1}{|\boldsymbol{A}|}\text{adj}\,\mathbf{A} = \frac{1}{8}\begin{bmatrix} 4 & -2 \\ -8 & 6 \end{bmatrix} = \begin{bmatrix} .50 & -.25 \\ -1.00 & .75 \end{bmatrix}$$

It was mentioned earlier that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. For the present example,

$$\mathbf{A}^{-1}\mathbf{A} = \begin{bmatrix} .50 & -.25 \\ -1.00 & .75 \end{bmatrix}\begin{bmatrix} 6 & 2 \\ 8 & 4 \end{bmatrix} = \begin{bmatrix} 1.00 & 0 \\ 0 & 1.00 \end{bmatrix}$$

$$\mathbf{A}^{-1} \qquad \mathbf{A} \qquad \mathbf{I}$$

It was said above that a matrix with a determinant of zero is singular. From the foregoing demonstration of the calculation of the inverse, it should be clear that a singular matrix has no inverse. Although one does not generally encounter singular matrices, an unwary researcher may introduce singularity in the treatment of the data. For example, suppose that a test battery consisting of five subtests is used to predict a given criterion. If, under such circumstances, the researcher uses not only the results on the five subtests but also a total test, obtained as the sum of the five subtests, he or she has introduced a linear dependency (see above), thereby rendering the matrix singular; similarly, when one uses tests on two scales as well as the differences between them in the same matrix. Other situations when one should be on guard not to introduce linear dependencies in a matrix occur when coded vectors are used to represent categorical variables.

It is realized that this brief introduction to matrix algebra cannot serve to demonstrate its great power and elegance. To do this, it would be necessary to use matrices with dimensions larger than the ones used here for simplicity of presentation. To begin to appreciate the power of matrix algebra, it is suggested that you think of the large data matrices frequently encountered in product development of any large corporation. Using matrix algebra, one can manipulate and operate upon large matrices with relative ease, when ordinary algebra will simply not do. For example, when in multiple regression analysis only two independent variables are used, it is relatively easy to do the calculations by ordinary algebra. But with increasing numbers of independent variables, the use of matrix algebra for the calculation of multiple regression analysis becomes a must. And in addition, matrix algebra is the language of linear structural equation models and multivariate analysis. In short, to understand and be able to intelligently apply these methods, it is essential that you develop a working knowledge of matrix algebra. It is therefore strongly suggested that you do a serious review of some of the references already mentioned in this appendix. Furthermore, it is suggested that you learn to use computer programs when you have to manipulate relatively large matrices. Of the various computer programs for matrix manipulations, one of the best and most versatile is the MATRIX program of the SAS package, followed by SPSS and Minitab.

REFERENCES

Dorf, R., *Matrix Algebra*, John Wiley & Sons, New York, 1969.

Green, P.E., *Mathematical Tools for Applied Multivariate Analysis*, Academic Press, New York, 1976.

Hohn, F.E., *Elementary Matrix Algebra*, MacMillan, New York, 1964.

Horst, P., *Matrix Algebra for Social Science*, Holt, Rinehart & Winston, New York, 1963.

Searle, S.R., *Matrix Algebra for the Biological Sciences,* John Wiley & Sons, New York, 1966.

Strang G., *Linear Algebra and Its Applications,* 2nd ed., Academic Press, New York, 1980.

Appendix B — The Simplex Method in Two Dimensions

Although the geometrical method for solving linear programming problems is highly satisfactory for problems that involve only two variables, it is a very difficult one to use on problems involving several variables. For those problems an algebraic method is needed. One of the most popular such algebraic devices is known as the simplex method. Although it was designed to solve more complicated problems, we will apply it to a two-dimensional problem in order to simplify the explanation.

The problem is to find the values of x and y that maximize the linear function

$$f = 3x + 2y \tag{1}$$

subject to the restrictions

$$\begin{aligned} 2x - y &\leq 1 \\ x + 2y &\leq 3 \\ x &\geq 0 \\ y &\geq 0 \end{aligned}$$

The solution to this problem can be obtained by inspecting the nature of the feasibility region as shown in Figure B.1 and realizing that f is maximized by a family of lines $3x + 2y = f$, which passes through the point (1,1).

It is true in general that a set of linear inequalities in two variables will always give rise to a feasibility region whose boundary consists of line segments. It is this property of the feasibility region that guarantees that a linear function f will take on its maximum (or minimum) value at a corner point of the boundary.

Although we are studying the problem here in two dimensions only, it can be shown by algebraic techniques that a similar property of the feasibility region is true

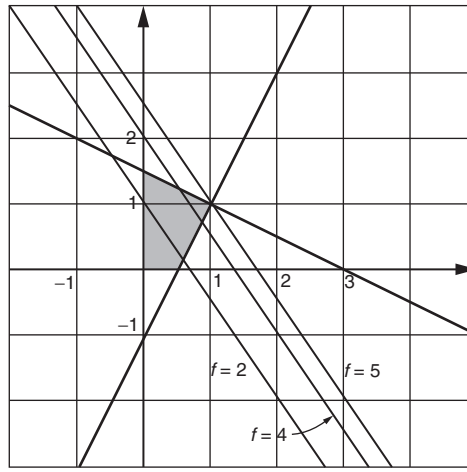


FIGURE B.1 A typical geometry of linear programming.

in higher dimensions. Thus a linear function of k variables subject to a set of linear inequalities in those variables will assume its maximum (or minimum) value at one of the corners of the feasibility region that is determined by those inequalities. In practice we assume this fact without proof.

In view of the preceding comments, it should suffice to find all the corner points of the feasibility region and then calculate the value of f at each of those points to determine which of them produces the maximum (or minimum) value of f. Although such a procedure is carried out for a two-dimensional problem, the same is not true for higher dimensional problems because the number of corners may become very large, and much algebra is needed to locate them. The simplex method is a method that allows us to start at any given corner point and then proceed step-by-step to a neighboring corner that yields a larger value of f until the maximum corner is reached. For a minimizing problem, successive corners are chosen that decrease the value of f each time. We restrict ourselves to maximizing problems, because a problem in which f is to be minimized can be converted to one in which g is to be maximized by setting g = –f.

With these preliminaries out of the way, we are ready to solve the problem stated in (1) by the simplex method. This method begins by introducing as many new variables as are needed to convert the inequalities into equalities, except for inequalities of the type $x \geq 0$ and $y \geq 0$. An inequality such as $2x - y \leq I$ can be made into an equality by introducing a new variable, say r, and writing

$$2x - y + r = 1$$

If x and y have values such that $2x - y < I$, then r will be a positive number which, when added to 2x – y, will make the sum equal 1. If x and y have values such that $2x - y = 1$, then r will have the value 0. Thus, r is a nonnegative number that transforms the inequality $2x - y \leq 1$ into the equality $2x - y + r = 1$. Similarly,

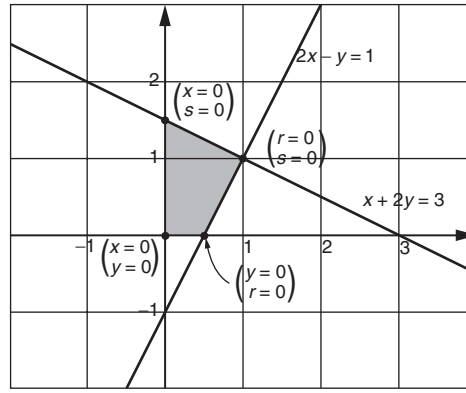


FIGURE B.2 The simplex notation for corner points.

we can introduce a nonnegative variable s that will transform the inequality x + 2y ≤ 3 into the equality x + 2y + s = 3. The variables r and s that have been introduced here are called slack variables because they take up the slack on the left side of an inequality of the form ≤ to make the equality sign hold. In this connection, the variables x and y are called the basic variables.

The linear programming problem of (1) may now be reformulated as the problem of maximizing the function

$$f = 3x + 2y$$

subject to the restrictions

$$\begin{gathered} 2x - y + r = I \\ x + 2y + s = 3 \\ x \geq 0, \quad y \geq 0, \quad r \geq 0, \quad s \geq 0 \end{gathered} \tag{2}$$

Next, we express the corner points of our feasibility region in terms of the values of the four variables x, y, r, and s. The feasibility region for this problem was obtained earlier and can be found in Figure B.1. It is duplicated in Figure B.2. Each of these corner points is the point of intersection of two lines out of the set:

$$x = 0, \quad y = 0, \quad 2x - y = 1, \quad \text{and} \quad x + 2y = 3$$

Since the last two lines of this set are special cases of the first two equalities in (2) and are characterized by setting r = 0 and s = 0, respectively, we may represent this set of lines by the symbols:

$$x = 0, \quad y = 0, \quad r = 0, \quad \text{and} \quad s = 0$$

The origin is a corner point and is labeled (x = 0, y = 0) because it is the intersection of the y axis (x = 0) and the x axis (y = 0). The corner point on the x

axis is labeled (y = 0, r = 0) because it is the intersection of the x axis (y = 0) and the line 2x – y = 1 (r = 0). The corner point labeled (r = 0, s = 0) is so labeled because it is the intersection of the two lines 2x – y = 1 (r = 0) and x + 2y = 3 (s = 0). Finally, the corner point on the y axis is labeled (x = 0, s = 0) because it is the intersection of the y axis (x = 0) and the line x + 2y = 3 (s = 0). The labeling of all the corner points of the feasibility region in this manner is shown in Figure B.2.

The next step in the simplex method is to start at a corner point and proceed to a neighboring corner that will increase the value of f, assuming that the maximum value of f has not already been attained. Suppose that we start at the (x = 0, y = 0) corner. Then we wish to proceed to the (y = 0, r = 0) corner or to the (x = 0, s = 0) corner. It should be observed that this procedure leaves one of x = 0 or y = 0 alone and replaces the other by the zero value of a different variable. Since our function f = 3x + 2y will grow faster if x is increased one unit than if y is increased one unit, and we wish to make f as large as possible, we agree to leave y = 0 alone and increase x as much as possible. But from Figure B.2 this means that we choose the (y = 0, r = 0) corner in preference to the (x = 0, s = 0) corner. Setting y = 0 and r = 0 in equations (2), we obtain

$$2x = 1$$
$$x + s = 3$$

Solving these equations, we find that x = 1/2, s = 5/2, and f = 3/2. Since f had the value 0 and our starting corner (x = 0, y = 0), this shift has increased its value from 0 to 3/2.

We now repeat the performance, beginning with the (y = 0, r = 0) corner and treating y and r as the basic variables and x and s as the slack variables. To do so, we must express f as a function of y and r only. This can be accomplished by expressing x as a function of y and r and substituting it into the expression for f. We therefore solve the first equation in (2) for x in terms of y and r and substitute it into the expression for f. Thus

$$f = 3\left(\frac{1 + y - r}{2}\right) + 2y = \frac{3}{2} + \frac{7y}{2} - \frac{3r}{2}$$

Now treating y and r as the basic variables, it is clear from this expression that f can be increased by increasing y from its zero value. Increasing r from its zero value, however, would decrease the value of f. This implies that we should hold r fixed at its zero value and should increase y as much as possible. Geometrically, this means that we should move from the (y = 0, r = 0) corner to the neighboring corner where r = 0 and y is positive, which from Figure B.2 is the (r = 0, s = 0) corner. Setting r = 0 and s = 0 in equations (2), we obtain

$$2x - y = 1$$
$$x + 2y = 3$$

Solving these equations, we get x = 1, y = 1, and f = 5. Since the value of f has increased from 3/2 to 5, this shift has increased the value of f further.

Although we know from our earlier result that we have reached the maximizing corner, we proceed as though we were not aware of this fact. Thus, since our new basic variables are chosen to be r and s, we must express f in terms of r and s. This is accomplished by solving equations (2) for x and y in terms of r and s and substituting those values into f. The solution of equations (2) is given by

$$x = 1 - 2r/5 - s/5$$
$$y = 1 + r/5 - 2s/5$$

As a result, f assumes the form

$$f = 3\left(1 - \frac{2r}{5} - \frac{s}{5}\right) + 2\left(1 + \frac{1}{5}r - \frac{2}{5}s\right) = 5 - \frac{4r}{5} - \frac{7s}{5}$$

Since r and s cannot be increased from their zero values without decreasing the value of f, it follows that r = 0 and s = 0 yield the maximum value of f, namely 5.

This technique of shifting to a neighboring corner that increases the value of f until the maximizing corner has been reached assumes that it is always possible to arrive at the maximizing corner in this manner. A justification for this assumption can be given that is based on the nature of the feasibility region. Since this property of our feasibility region seems obvious from the geometry of the problem for two-dimensional problems, we do not attempt a justification. The chief advantage of our method is that it eliminates the necessity of finding the coordinates of all the corner points and of checking the value of f at each of them. As we stated previously, the latter method can become a lengthy computational problem in higher dimensions.

Another striking advantage of the simplex method is that the technique can be carried out in a systematic routine manner by means of matrix methods, regardless of the number of variables involved. It is not necessary to perform any of the geometry that was used in our present problem. The geometry was introduced to explain how and why the simplex method works.

Appendix C — Bernoulli Trials

In this appendix, we consider a particular random variable that is a generalization of the coin-tossing random variable. Toward this objective, consider an experiment in which the outcome can always be classified as a success or a failure. In the coin-tossing experiment, success would correspond to getting a head, and failure would correspond to getting a tail. Let p denote the probability of getting a success and let q = 1 – p denote the failure probability, Further, let the experiment be repeated n times, and let x denote the total number of successes that will be obtained in the n repetitions of the experiment. In the coin-tossing experiment, for example, we would have p = q = 1/2 and n = 3. In terms of this notation the basic problem is to find the probability distribution of the random variable x.

It is assumed in problems of this type that the n experiments are independent in a probability sense and, therefore, that the multiplication rule for independent events may be applied to them. The n independent repetitions of the experiment are usually called the n trials of the experiment. A sequence of independent trials such as this in which the probability of success is the same for all trials is called a sequence of Bernoulli trials. The name is in honor of a Swiss mathematician who pioneered in the study of probability. The coin-tossing experiment is an illustration of a sequence of three Bernoulli trials for which the probability of success in any given trial is 1/2.

The technique for finding the probability distribution of x is a generalization of that used in Chapter 16 for the coin-tossing problem. We first calculate the probabilities for all possible sequences of outcomes and then add the probabilities of those sequences that yield the same value of x. Suppose, for example, that we wish to calculate P{k}, where this symbol denotes the probability that the random variable x will assume the value k, and where k is some integer between 0 and n. One possible sequence that will make x = k is the following one in which all the successes occur first, followed by all failures:

$$\overbrace{SS\ldots S}^{k}\ \overbrace{FF\ldots F}^{n-k}$$

Another such sequence is the following one in which a failure occurs first, followed by k consecutive successes, then followed by the remaining failures. Thus

$$\overbrace{FSS\ldots S}^{k}\ \overbrace{FF\ldots F}^{n-k-1}$$

Because of the independence of the trials, the probability of obtaining the first of these two sequences is given by

$$\overbrace{p\cdot p\ldots p}^{k}\ \overbrace{qq\ldots q}^{n-k}=p^k q^{n-k}$$

The probability for the second sequence is given by

$$q\cdot\overbrace{p\cdot p\ldots p}^{k}\cdot\overbrace{q\cdot q\ldots q}^{n-k-1}=p^k q^{n-k}$$

The probability for the two sequences is the same and clearly will be the same for every sequence that satisfies the condition of having k successes and n – k failures.

The number of ways in which the desired event can occur is equal to the number of different sequences that can be written down of the type just displayed, those containing k letters S and n – k letters F. But this number is equal to the number of ways of choosing k positions out of n positions along a line in which to place the letter S. The remaining n – k positions will automatically be assigned the letter F. Since we are interested only in which of the n positions are to be selected and not in the order in which we choose them, this is a combination problem of choosing k things from n things. From formula

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

now we can determine the number of such sequences.

Since each of these sequences represents one of the mutually exclusive ways in which the desired event can occur, and each such sequence has the same probability of occurring, namely $p^k q^{n-k}$, it follows that the desired probability is obtained by adding this probability as many times as there are sequences. But the number of sequences was just found to be $\binom{n}{k}$; therefore, P{k} is obtained by multiplying $p^k q^{n-k}$ by $\binom{n}{k}$. Hence:

$$P\{k\} = \frac{n!}{k!(n-k)!} p^k q^{n-k} \qquad k = 0,1,\ldots,n$$

This formula gives the probability of x = k, that is, of obtaining k successes in n Bernoulli trials for which the probability of success in a single trial is p. The random variable x is commonly called the binomial variable, and the above formula is a formula for the binomial distribution.

Although the problems used to motivate this derivation have been related here to games of chance, there are many types of practical problems that can be solved by means of the binomial distribution formula. We consider only a few simple problems that require little computation to illustrate its use.

Example 1

The probability that parents with a certain type of blue-brown eyes will have a child with blue eyes is 1/4. If there are six children in the family, what is the probability that at least half of them will have blue eyes? To solve this problem, the six children in the family will be treated as six independent trials of an experiment for which the probability of success in a single trial is 1/4. Thus n = 6 and p = 1/4 here. It is necessary to calculate P{3}, P{4}, P{5}, and P{6} and sum them because these probabilities correspond to the mutually exclusive ways in which the desired event can occur. By the use of the binomial distribution formula,

$$P(3) = \frac{6!}{3!3!}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^3 = \frac{540}{4096}$$

$$P(4) = \frac{6!}{4!2!}\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^2 = \frac{135}{4096}$$

$$P(5) = \frac{6!}{5!1!}\left(\frac{1}{4}\right)^5\left(\frac{3}{4}\right)^1 = \frac{18}{4096}$$

$$P(6) = \frac{6!}{6!0!}\left(\frac{1}{4}\right)^6\left(\frac{3}{4}\right)^0 = \frac{1}{4096}$$

The probability of getting at least three successe is obtained by adding these probabilities; consequently, by writing $x \geq 3$ to represent at least three successes, we obtain

$$\mathrm{P}\{x \geq 3\} = \frac{694}{4096} = .169$$

This result shows that there is a very small chance that a family such as this will have so many blue-eyed children. In only about 17 of 100 such families will at least half of the children be blue-eyed.

Example 2

A manufacturer of certain parts for automobiles guarantees that a box of the company's parts will contain at most two defective items. If the box holds 20 parts and experience has shown that the manufacturing process produces 2% defective items, what is the probability that a box of the parts will satisfy the guarantee? This problem can be considered as a binomial distribution problem for which n = 20 and p = .02. A box will satisfy the guarantee if the number of defective parts is 0, 1, or 2. By means of the binomial distribution formula the probabilities of these three events are given by

$$P\{0\} = \frac{20!}{0!20!}(.02)^0(.98)^{20} = (.98)^{20} = .668$$

$$P\{1\} = \frac{20!}{1!19!}(.02)^1(.98)^{19} = 20(.02)(.98)^{19} = .273$$

$$P\{2\} = \frac{20!}{2!18!}(.02)^2(.98)^{18} = 190(.02)^2(.98)^{18} = .053$$

The calculations here were made with the aid of logarithms. Since these are mutually exclusive events, the probability that there will be at most two defective parts, written $x \leq 2$, is the sum of these probabilities; hence, the desired answer is

$$P\{x \leq 2\} = .994$$

This result shows that the manufacturer's guarantee will almost always be satisfied.

Example 3

As a final illustration, consider the following problem concerning whether it pays to guess on an examination. Suppose an examination consists of 10 questions of the multiple-choice type, with each question having five possible answers but only one of the five being the correct answer. If a student receives 3 points for each correct answer and –1 point for each incorrect answer, and if on each of the 10 questions his probability of guessing the correct answer is only 1/3, what is the student's probability of obtaining a positive total score on those 10 questions?

If x denotes the number of questions answered correctly, then a positive score will result if $3x > 10 - x$ because the left side of this inequality gives the total number of positive points scored and the right side gives the total number of penalty points. This inequality will be satisfied if $x > 10/4$, which implies that at least three correct answers must be obtained to realize a positive score. The desired probability is therefore given by

$$P\{x \geq 3\} = 1 - \sum_{x=0}^{2} \frac{10!}{x!(10-x)!}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{10-x}$$

$$= 1 - \left\{\left(\frac{2}{3}\right)^{10} + 10\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^9 + 45\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^8\right\}$$

$$= .70$$

Thus the student has an excellent chance of gaining points if his or her probability of guessing a correct answer is as high as 1/3. If the student knew nothing about the material and selected one of the five alternatives by chance, the probability would, of course, be only 1/5 for each question. It is assumed here, however, that the student knows enough about the subject to be able to discard two of the five possibilities as being obviously incorrect and to make a guess regarding the other three. If the student had no such knowledge, so that his or her probability would be 1/5, then similar calculations would show that it would not pay to guess.

Appendix D — Markov Chains

In this appendix, we construct a probability model for a sequence of events that is a generalization of Bernoulli trials. Recall that Bernoulli trials are independent events and that the probability of success is constant from trial to trial. We generalize this model in two ways. First, we permit the number of possible outcomes to be some positive integer k where $k \geq 2$. Thus we no longer can speak only of success or failure at a trial. Let $A_1,A_2,...,A_k$ denote those possible outcomes. Second, we drop the assumption that the trials are independent events and introduce a certain amount of dependence. A large share of the interesting random variables in the social sciences as well as in the financial and engineering worlds are variables that are observed at regular time intervals over a given period of time. Very often those variables are not independent. For example, the price of a given individual stock may vary from week to week in a random manner, but its price during one week will usually depend rather heavily on what its price was during the preceding week. As another illustration, if a random variable represents the number of items, out of a set of n learned items, that will be recalled after, say, t time intervals have elapsed, then the value of that variable will certainly depend on how many items were recalled after $t - 1$ time intervals.

In some problems of the preceding type, the dependence is a local one in the sense that the value of the random variable at the end of t time intervals depends on its value at the end of $t - 1$ time intervals but not on any earlier time interval values. This would not usually be true, however, for the stock market because many buyers look at the past performance of a stock, and not merely at its price last week, in determining whether to buy it this week. In this section, we consider this special model in which the dependence of a random variable on earlier random variables extends only to the immediately preceding one. Thus we assume that the probability of outcome A_j occurring at a given trial depends on what outcome occurred at the immediately preceding trial but on no others.

Let P_{ij} denote the probability that outcome A_j will occur at a given trial if outcome A_i occurred at the immediately preceding trial. The possible outcomes $A_1,A_2,...,A_k$ are called the states of the system. Thus P_{ij} is the probability of going from state A_i

to state A_j at the next trial of the sequence. A sequence of experiments of the preceding type is called a Markov sequence or a Markov chain.

It should be noted that a Bernoulli sequence is a special case of a Markov sequence in which there are only two states A_1 and A_2 corresponding to success and failure and in which $P_{11} = p$, $P_{12} = q$, $P_{21} = p$, and $P_{22} = q$.

The probabilities P_{ij}, which are called transition probabilities, are usually displayed in matrix form as follows:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix}$$

These probabilities apply to the system at any point in time. They express the probability relationship that exists at two neighboring time points, regardless of the chosen point in time.

The preceding probabilities are one-step transition probabilities. There are, however, also two-step and more generally n-step transition probabilities. A two-step transition probability, denoted by $P_{ij}^{(2)}$, is the probability that the system will be in state A_j two time intervals later if it is now in state A_i. Formulas for expressing multiple-step transition probabilities in terms of one-step probabilities can be obtained by using matrix methods and the rules of probability. For example, to obtain a formula for $P_{ij}^{(2)}$, we proceed as follows.

To arrive at A_j from A_i in exactly two steps, we must first go from A_j to A_r, where r is one of the integers 1,2,…,k and then go from A_r to A_j at the next step. The probability of doing this is $P_{ir}P_{rj}$. Since there are k mutually exclusive ways in which the desired event can occur corresponding to r = 1,2,…,k, the sum of those probabilities must be the value of $P_{ij}^{(2)}$, Hence, we have the formula

$$P_{ij}^{(2)} = \sum_{r=1}^{k} P_{ir}P_{rj} \tag{1}$$

Now consider the evaluation of P^2, that is, of the matrix product

$$P^2 \begin{bmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & & \vdots \\ p_{i1} & \cdots & p_{ik} \\ \vdots & & \vdots \\ p_{k1} & \cdots & p_{kk} \end{bmatrix} \begin{bmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1k} \\ \vdots & & \vdots & & \vdots \\ p_{k1} & \cdots & p_{kj} & \cdots & p_{kk} \end{bmatrix}$$

To obtain the element in the ith row and jth column of P^2, it is necessary to multiply the ith row vector in the first matrix by the jth column vector in the second matrix. This gives the vector product

$$P_{i1}P_{1j} + P_{i2}P_{2j} + \ldots + P_{ik}P_{kj}$$

But this is the same as the expression defining $P_{ij}^{(2)}$ in (1); hence, it follows that P^2 is the matrix whose typical element is $P_{ij}^{(2)}$.We may express this relationship by writing

$$P^2 = [P_{ij}^{(2)}] \tag{2}$$

This shows that the two-step transition probabilities may be obtained by merely multiplying the one-step transition probability matrix by itself. In a similar manner, it follows that the n-step transition probabilities, $p_{ij}^{(n)}$, are given by the relationship

$$p^n = [p_{ij}^{(n)}]$$

As an illustration of a Markov chain, we consider a problem related to politics. Suppose that of the sons of Republican fathers, 60 percent vote Republican, 30 percent vote Democratic, and 10 percent vote Socialist; of the sons of Democratic fathers, 60 percent vote Democratic, 20 percent vote Republican, and 20 percent vote Socialist; of the sons of Socialist fathers, 50 percent vote Socialist, 40 percent vote Democratic, and 10 percent vote Republican. With this information, and assuming that the Markov chain properties hold true, we wish to perform the following:

1. Write down the transition matrix.
2. Calculate the two-step transition matrix.
3. Determine the probability that the grandson of a Republican man will vote Democratic.
4. Determine the same probability for a great-grandson.

Example:

1. Using the order Republican, Democrat, Socialist, the transition matrix is

$$P = \begin{bmatrix} .6 & .3 & .1 \\ .2 & .6 & .2 \\ .1 & .4 & .5 \end{bmatrix}$$

2. By using formula (2)

$$P^2 = \begin{bmatrix} .6 & .3 & .1 \\ .2 & .6 & .2 \\ .1 & .4 & .5 \end{bmatrix} \begin{bmatrix} .6 & .3 & .1 \\ .2 & .6 & .2 \\ .1 & .4 & .5 \end{bmatrix}$$

$$= \begin{bmatrix} .43 & .40 & .17 \\ .26 & .50 & .24 \\ .19 & .47 & .34 \end{bmatrix}$$

3. The answer here is given by $P_{12}^{(2)}$; hence, picking out the element in the first row and second column of P^2,we obtain $P_{12}^{(2)} = .40$.
4. Here we need the element $P_{12}^{(3)}$, which can be obtained by multiplying the first row vector of P by the second column vector of P^2. This gives

$$(.6)(.40) + (.3)(.50) + (.1)(.47) = .437$$

In the preceding problem, suppose that the proportions of Republicans, Democrats, and Socialists in a community are given by the row vector $\mathbf{A} = [a_1\ a_2\ a_3]$, where $a_1 + a_2 + a_3 = 1$. Then the matrix product

$$\boldsymbol{AP} = [a_1 \quad a_2 \quad a_3]\begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

is a row vector whose components give the proportions of Republicans, Democrats, and Socialists in the first generation of offspring (sons). For example, the first component in this product, namely

$$a_1P_{11} + a_2P_{21} + a_3P_{31}$$

gives the probability that a first generation offspring will be a Republican, because it is the sum of the probabilities of the mutually exclusive ways in which a son will become a Republican. He may start with a Republican father and then become a Republican; or he may start with a Democratic father and then become Republican; or he may start with a Socialist father and then become a Republican. This works similarly for the other two components. In a similar manner, if we wish to calculate the proportions of male offspring who will be Republicans, Democrats, or Socialists at the nth generation, we merely need to calculate the matrix product $\mathbf{AP}^n$.

As an illustration, suppose that the preceding initial proportions are given by the row vector $\mathbf{A} = [.4\ .5\ .1]$. Then after one generation the proportions become

$$\boldsymbol{AP} = [.4 \quad .5 \quad .1]\begin{bmatrix} .6 & .3 & .1 \\ .2 & .6 & .2 \\ .1 & .4 & .5 \end{bmatrix}$$

$$= [.35 \quad .46 \quad .19]$$

After two generations, the proportions become

$$\boldsymbol{AP}^2 = [.4 \quad .5 \quad .1]\begin{bmatrix} .43 & .40 & .17 \\ .26 & .50 & .24 \\ .19 & .47 & .34 \end{bmatrix}$$

$$= [.321 \quad .457 \quad .222]$$

It would be interesting to carry these calculations further for succeeding generations to observe whether these proportions approach some fixed values. Calculating high powers of a matrix requires high-speed computing equipment and may be very expensive. Fortunately, however, there exists a mathematical theorem that tells us what happens to the transition probabilities in a Markov chain as n becomes increasingly large, without the necessity of extensive calculations. This theorem, which we do not prove, can be expressed as follows:

Theorem. As $n \to \infty$ each row vector of P_n approaches the probability vector X that is a solution of the matrix equation XP = X.

It is easy to show that if P is a transition probability matrix, which implies that its elements are nonnegative and that the sum of the elements in any row is 1, then P^n will also be a matrix of this type. Our theorem therefore states that the resulting transition probabilities are the same for all rows of the matrix.

As an illustration, we compute this transition matrix, which is often called the limiting transition matrix, for the preceding political problem. The equation that needs to be solved is the following one:

$$[x_1 \quad x_2 \quad x_3]\begin{bmatrix} .6 & .3 & .1 \\ .2 & .6 & .2 \\ .1 & .4 & .5 \end{bmatrix} = [x_1 \quad x_2 \quad x_3]$$

Multiplying the matrices on the left and equating components on both sides we obtain

$$\begin{aligned} .6x_1 + .2x_2 + .1x_3 &= x_1 \\ .3x_1 + .6x_2 + .4x_3 &= x_2 \\ .1x_1 + .2x_2 + .5x_3 &= x_3 \end{aligned}$$

These equations are equivalent to the set

$$\begin{aligned} -.4x_1 + .2x_2 + .1x_3 &= 0 \\ .3x_1 - .4x_2 + .4x_3 &= 0 \\ .1x_1 + .2x_2 - .5x_3 &= 0 \end{aligned}$$

They simplify to

$$\begin{aligned} -4x_1 + 2x_2 + x_3 &= 0 \\ 3x_1 - 4x_2 + 4x_3 &= 0 \\ x_1 + 2x_2 - 5x_3 &= 0 \end{aligned}$$

Since the solution to our problem requires that X be a probability vector, we must have $x_1 + x_2 + x_3 = 1$. The solution of the preceding equations that also satisfies this restriction is readily seen to be given by

$$x_1 = \frac{12}{41}, \quad x_2 = \frac{19}{41}, \quad x_3 = \frac{10}{41}$$

Our theorem then states that as $n \to \infty$, the elements of P^n approach the elements of the matrix

$$\begin{bmatrix} \frac{12}{41} & \frac{19}{41} & \frac{10}{41} \\ \frac{12}{41} & \frac{19}{41} & \frac{10}{41} \\ \frac{12}{41} & \frac{19}{41} & \frac{10}{41} \end{bmatrix}$$

Now let us calculate the long-run proportions of Republicans, Democrats, and Socialists. Just as was done earlier to obtain the first- and second-generation proportions, this can be accomplished by calculating the product of the row vector [.4 .5 .1] and the matrix just obtained. This gives

$$[.4 \quad .5 \quad .1]\begin{bmatrix} \frac{12}{41} & \frac{19}{41} & \frac{10}{41} \\ \frac{12}{41} & \frac{19}{41} & \frac{10}{41} \\ \frac{12}{41} & \frac{19}{41} & \frac{10}{41} \end{bmatrix} = \begin{bmatrix} \frac{12}{41} & \frac{19}{41} & \frac{10}{41} \end{bmatrix}$$

It should be observed that since the rows of the limiting transition matrix are all equal and the elements of the row vector A sum to 1, it follows that the result of this type of matrix product must be a row of the limiting transition matrix. Hence, the preceding solution

$$x_1 = \frac{12}{41}, \quad x_2 = \frac{19}{41}, \quad x_3 = \frac{10}{41}$$

gives the proportions of Republicans, Democrats, and Socialists that will be realized in the long run if the transition probabilities do not change over time. Incidentally, this also shows that the initial set of proportions, .4, .5, and .1, has no effect on the long-run set of proportions.

The preceding result, showing that the limiting proportions are the same as the limiting values of the transition probabilities, is quite general. It did not depend on the particular numbers used in the illustration. Thus it is true in general that the initial proportions in any problem have no effect on the long-run proportions. Those proportions depend only on the nature of the transition probabilities.

Appendix E — Optimization

Consider the following simple 2×3 payoff matrix in which the elements represent dollar rewards to R.

$$\begin{bmatrix} 5 & 2 & 3 \\ -5 & 1 & 6 \end{bmatrix}$$

We wish to determine which strategies R and C should choose. Suppose that R chooses R_1, which means that he chooses the first row, then he can be certain of winning at least \$2 and at most \$5. If he chooses R_2, he could lose \$5 but he might win as much as \$6. In view of the wide range of possibilites here, it appears that R should examine more carefully his possible strategies to determine how much he can win even if C is clever or lucky enough to choose his best possible strategy each time.

If R were to choose R_1, C would certainly choose C_2 and thus would limit R to winning \$2. If R were to choose R_2, C would obviously choose C_1 and thus would cause R to lose \$5. The two elements that represent these two possibilites have been enclosed in circles in the matrix of Figure E.1. They represent the worst that can happen to R regardless of how clever C may be in anticipating which strategy R will select. If R wishes to protect himself as best possible against these undesirable possibilities, he should choose R_1. Choosing R_1 makes it certain that R will win at least \$2, and possibly more if it should happen that C does not select his best strategy. Although neither opponent knows which strategy the other will choose, one of them may outguess the other and may gain considerably thereby. The preceding analysis guards R against the occurrence of this possibility.

Now consider which strategy C should employ. He wishes to keep the amount to be won by R as small as possible because he is required to pay R when R wins. If he reasons in the same manner as R does, he will examine each of his possible strategies to see how well he can do even if R is clever enough to anticipate which strategy C will select. Thus, if C were to select C_1, R would choose R_1 and win \$5.

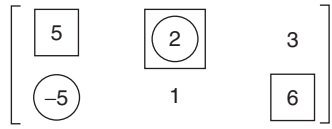


FIGURE E.1 Strategies for R and C.

If C were to select C_2, R would choose R_1 and win \$2. If C were to select C_3, R would choose R_2 and win \$6. Each of the elements representing these three possibilities has been enclosed in a square in Figure E.1. They represent the worst that can happen to C regardless of how clever or lucky R may be. If C wishes to protect himself as best possible against these undesirable possibilities, he should choose C_2 because that minimizes the amount that he will need to pay to R.

The preceding analysis, which assumes that R wishes to maximize his winnings and C wishes to minimize his losses, leads to the conclusion that R should choose strategy R_1 and C should choose strategy C_2. If the game is played with these two strategies, R will win \$2 from C. The number 2 is called the value of this game. It represents the amount that R expects to win and C expects to pay if R and C play their best defensive strategies.

If the best strategies for R and C do not yield a common circle-square element of the payoff matrix, the game does not have a solution, and therefore does not have a value.

Appendix F — Randomized Strategies

The game described in Appendix E was played under the assumption that it was to be played only once. Hence, once R and C have selected their strategies, the result of the game is determined. Suppose, however, that a sequence of games is to be played and that R and C are permitted to change their strategies from game to game. How should they proceed then? Since each contestant assumes that the other is just as intelligent as he is, they dare not develop a systematic pattern of choosing strategies for fear of this being discovered; hence, it would be advisable for them to choose their successive strategies by means of some random scheme. This can be accomplished by having R select a sect of probabilities $p_1, p_2, \ldots, p_m$ that will determine the relative frequencies with which he wishes his strategies $R_1, R_2, \ldots, R_m$ to be played. Similarly, C is permitted to select a set of probabilities $q_1, q_2, \ldots, q_n$ that will determine the relative frequencies with which he wishes his strategies $C_1, C_2, \ldots, C_n$ to be played.

As an illustration, in the game of the preceding section R might choose

$$p_1 = \frac{2}{3} \text{ and } p_2 = \frac{1}{3}, \text{ and C might choose } q_1 = \frac{2}{4}, \; q_2 = \frac{1}{4}, \text{ and } q_3 = \frac{1}{4}$$

Each time the game is to be played, R will use a game of chance that yields p_1 twice as frequently as p_2 to choose one of those strategies. This could be done, for example, by drawing a card from a set of three cards that contains two aces and one deuce. Similarly, C could draw a card from a set of four cards that contains two aces, one deuce, and one three.

Strategies that are selected by chance according to a set of probabilities are called randomized strategies. They include the one-play strategies that were obtained in the preceding section because it is merely necessary to choose $p_1 = 1$, $p_2 = 0$, and $q_l = 0$, $q_2 = 1$, $q_3 = 0$ to arrive at the strategies R_1 and C_2 that were determined for that game. If a player uses a set of probabilities in which one of them is 1 and all the rest are 0, he is said to be using a pure strategy, otherwise he is using a mixed strategy.

Even though a game is to be played only once, and this is the natural situation in most business-type games, it may be that one of the competitors can do better than the other by employing randomization in choosing his strategy. Therefore, we take a fresh look at our earlier one-play games in studying randomized games to see if improvements are possible by using randomization.

Since the payoffs will vary from game to game because they will depend on chance, we look at the average payoff in a long sequence of games. This is equivalent to looking at the expected value of the payoff. Now the expected value of the payoff in the ith row and jth column of the payoff matrix **A** is merely $a_{ij}P_iq_j$ because the row and column choices are made independently, and therefore the probability of R winning the amount a_{ij} is the product of the ith row and jth column probabilities.

As an illustration of a randomized payoff matrix, consider a matrix with the possibilities of $p_1 = \frac{2}{3}$, $p_2 = \frac{1}{3}$, $q_1 = \frac{1}{2}$, $q_2 = \frac{1}{4}$, and $q_3 = \frac{1}{4}$. Letting **B** denote this matrix we find that

$$B = \begin{bmatrix} 5 \cdot \frac{2}{3} \cdot \frac{1}{2} & 2 \cdot \frac{2}{3} \cdot \frac{1}{4} & 3 \cdot \frac{2}{3} \cdot \frac{1}{4} \\ -5 \cdot \frac{1}{3} \cdot \frac{1}{2} & 1 \cdot \frac{1}{3} \cdot \frac{1}{4} & 6 \cdot \frac{1}{3} \cdot \frac{1}{4} \end{bmatrix}$$

The expected payoff to R is the sum of these individual expected payoff values. Their sum is found to be 2 1/4; therefore, R would do slightly better under these two sets of randomized strategies than under the original nonrandomized version of the game. It may well be, however, that C chose a poor set of probabilities here. The purpose of this example is to illustrate how expected payoffs are calculated; it is not intended to illustrate good randomized strategies.

After a set of probabilities has been selected by each of R and C and the corresponding expected payoff matrix is calculated, the game is completely determined and could be played by a machine that selects successive pairs of strategies according to the probabilities $p_1,\ldots,p_m$ and $q_1,\ldots,q_n$.

The interesting question now is: How should R and C choose their probabilities? For example, is it possible for R to do better than C in the preceding illustrative game if his probabilities are chosen properly, regardless of what probabilities C chooses? In the preceding illustration, R did do better but perhaps C could have prevented this with a better set of probabilities. The answer to this question is as follows: Independent of what probabilities R selects, C can find a set of probabilities such that the expected value of the payoff to R will not exceed $2. In addition, regardless of what probabilities C may select, R can find a set of probabilities such that he can be assured of averaging at least $2. Thus neither R nor C can gain in this particular problem by using mixed strategies rather than pure strategies, provided that both R and C employ their best defensive randomized strategies.

Appendix G — Lagrange Multipliers

Problems of maximizing or minimizing a function of several variables when those variables satisfy some restriction equations can become quite difficult to solve. There is a technique, called the method of Lagrange multipliers, which often simplifies the calculations. It is named after the eighteenth-century French mathematician who introduced it.

For the purpose of illustrating the technique, we solve a simple problem involving only two variables. We wish to find where the function $f(x,y) = x^2 + 2xy$ assumes its minimum value if the variables are subject to the restriction equation $x^2y = 27$. This problem can be solved by solving for y in terms of x in the restriction equation and substituting it into f(x,y) to reduce f to a function of a single variable.

In the Lagrange multiplier method a new function is introduced in the following manner. First, we rewrite the restriction equation so that it assumes the form $g(x,y) = 0$. Hence, we write

$$g(x,y) = x^2y - 27 = 0$$

The new function then is the function defined by

$$\begin{aligned} F(x,y) &= f(x,y) - \lambda g(x,y) \\ &= x^2 + 2xy - \lambda(x^2y - 27) \end{aligned}$$

The parameter λ (lambda) is the Lagrange multiplier here. It always multiplies the restriction function after the restriction equation has been expressed in the form $g(x,y) = 0$. We now treat F(x,y) as though it were a function of x and y without any restriction on those variables. We therefore proceed to find where the function F(x,y) assumes its minimum value in the manner of the preceding sections. That is, we calculate F_x and F_y, set them equal to zero, find the critical points, and check to see which of those points, if any, yields the minimum. In this problem, we obtain

$$F_x = 2x + 2y - 2\lambda xy = 0$$

$$F_y = 2x - \lambda x^2 = 0$$

Since we have the restriction equation

$$x^2y - 27 = 0$$

in addition to the preceding two equations, we have three equations in the three unknowns: x, y, and lambda (λ). We proceed to solve them. Since we are not interested in the value of λ, we try to eliminate it first. Solving for λ in the second of the two partial derivative equations, we obtain

$$\lambda = \frac{2}{x}$$

This is substituted into the first of those two equations to give

$$2x + 2y - 4y = 0$$

which reduces to

$$x = y$$

The preceding three equations in three unknowns are now reduced to the following two equations in two unknowns:

$$x = y$$
$$x^2y - 27 = 0$$

The solution of this pair of equations is x = 3 and y = 3. As previously, it is easily shown that this pair of numbers minimizes $f(x,y) = x^2 + 2xy$, subject to the restriction $x^2y = 27$.

The reasoning behind the Lagrange multiplier technique is the following one. Suppose that we have found a point (x_0,y_0) that minimizes f(x,y) subject to the restriction g(x,y) = 0. Then this point must also minimize the function F(x,y) regardless of the value of λ because as long as the restriction g(x,y) = 0 is satisfied, the term λg(x,y) has the value zero and, hence, minimizing f(x,y) minimizes F(x,y). Conversely, any point (x_0,y_0) that minimizes F(x,y) and that also satisfies the restriction equation must minimize f(x,y) subject to this restriction because once more the term λg(x,y) has the value zero for any such point. Therefore if $F(x_0,y_0)$ is a minimum, $f(x_0,y_0)$ must be a minimum. Although this shows that minimizing F is equivalent to minimizing f, it does not prove that we may treat F as a function of x and y without any restrictions on those variables. A proof of this fact is rather involved and, therefore, will not be given here.

The same type of reasoning applies to problems involving functions of several variables and with several restrictions. We demonstrate the technique on the following problem.

Given the function f(x,y,z) = xyz, subject to the restriction xy + 2xz + 2yz = 24, find where it is maximized. We first write the restriction in the form of g(x,y,z) = xy + 2xz + 2yz – 24 = 0. Then we write

$$F(x, y, z) = f(x, y, z) - \lambda g(x, y, z)$$

$$= xyz - \lambda[xy + 2xz + 2yz - 24]$$

Next, we calculate the three partial derivatives and set them equal to zero. Thus

$$F_x = yz - \lambda[y + 2z] = 0$$

$$F_y = xz - \lambda[x + 2z] = 0$$

$$F_z = xy - \lambda[2x + 2y] = 0$$

To solve these equations we solve for λ in the last equation and substitute it into the first two

$$yz = [y + 2z]\frac{xy}{2x + 2y}$$

and

$$xz = [x + 2z]\frac{xy}{2x + 2y}$$

Multiplying through by 2x + 2y and collecting terms, we obtain

$$2y^2z = xy^2$$

and

$$2x^2z = x^2y$$

These equations can be written in factored form as

$$y^2(x - 2z) = 0$$

and

$$x^2(y - 2z) = 0$$

From the first equation, we obtain y = 0 and x = 2z. From the second equation we obtain x = 0 and y = 2z. The possible solutions then consist of the pairs

$$(y = 0, x = 0), \quad (y = 0, y = 2z), \quad (x = 2z, x = 0), \quad (x = 2z, y = 2z)$$

These are equivalent to

$$(y = 0, x = 0), \quad (y = 0, z = 0), \quad (x = 0, z = 0), \quad (x = 2z, y = 2z)$$

We still have the restriction equation to be satisfied, which is

$$xy + 2xz + 2yz = 24$$

The first of the four pairs of partial solutions does not satisfy the restriction equation; hence, it may be discarded. The second and third pairs also fail to satisfy the restriction equation, and they may also be discarded. Hence, we are left with the fourth pair only. Substituting those values into the restriction equation, we obtain

$$4z^2 + 4z^2 + 4z^2 = 24$$

This gives $z^2 = 2$ and $z = \pm\sqrt{2}$. There are therefore two legitimate critical points:

$$(2\sqrt{2}, 2\sqrt{2}, \sqrt{2}) \quad \text{and} \quad (-2\sqrt{2}, -2\sqrt{2}, -\sqrt{2})$$

The values of f(x,y,z) = xyz at these two points are $8\sqrt{2}$ and $-8\sqrt{2}$ respectively. Since we are interested in maximizing f we are left with the point $x = 2\sqrt{2}$, $y = 2\sqrt{2}$, $z = \sqrt{2}$.

This problem grew out of maximizing the volume of a box subject to the restriction that it could not use more than 24 square feet of material. We know from practical considerations that a finite maximum exists; therefore, this solution must produce an absolute maximum.

COMMENT

The problem of determining where a function of several variables assumes its maximum or minimum value is obviously an important one in all branches of science. The problems that we solved in this appendix are the basic kind of such problems. From these it is possible to branch out in many directions. There are many theoretical problems, for example, in economic theory and engineering as well as in many science situations in which there are several functions of several variables with functional relations and restrictions imposed on them, and the problem is to

maximize one of those functions subject to the restrictions. This is similar to the problem of linear programming, only here the functions are not linear.

The reader of this volume should be aware that a larger share of the applications of statistics and probability are concerned with how to maximize or minimize some function with some significant confidence. This obviously is a motivating force in calculus, but it is also the rationale for linear programming. Mathematical methods are certainly a powerful tool for solving this type of problem and would merit study for this reason only.

Appendix H — Monte Carlo Simulation

Decisions approaches include pure intuition and judgment, experience and analogy with similar situations, analysis with the aid of analytical models, experimentation with real systems, and experimentation with a model of a real system.

Simulation is associated with the last approach. It uses a model of a system and manipulates it so as to imitate the system's behavior over time for the purpose of evaluating alternative design characteristics or decision rules. It is a systematic trial-and-error method for solving complex problems. Simulation makes available an experimental laboratory for the experimenter by making it possible to test various alternatives without risking or committing organizational resources. The effects of numerous alternative policies can be ascertained without tampering with the actual system. This form of system experimentation can reduce the risk of upsetting the existing structure with changes that would not be beneficial. Simulation gives the manager an opportunity to test and evaluate proposals without running the risk of actually installing new approaches and absorbing the costs associated with the changes. With simulation, "trial and error" need not become "trial and catastrophe."

When problems involve risk or uncertainty, an analytical solution may be difficult or impossible to obtain. Simulation is useful in situations where analytical solutions are not appropriate because the models are either too complex or too costly. A mathematical model using the analytical approach can become incredibly complex because of numerous interacting variables. Simulation offers an alternative for complex problems not suitable for rigorous analytical analysis.

Simulation develops a model of some phenomenon and then performs experiments on the model. It is a descriptive rather than an optimization technique, which means that it does not yield optimal solutions. Monte Carlo simulation is a numerical technique that models a probabilistic system with the intention of predicting the system's behavior.

Monte Carlo simulation involves determining the probability distributions of the variables under study and then sampling from the distributions by using random numbers to obtain data. It is a probabilistic type of simulation that approximates the solution to a problem by sampling from a random process. A series of random

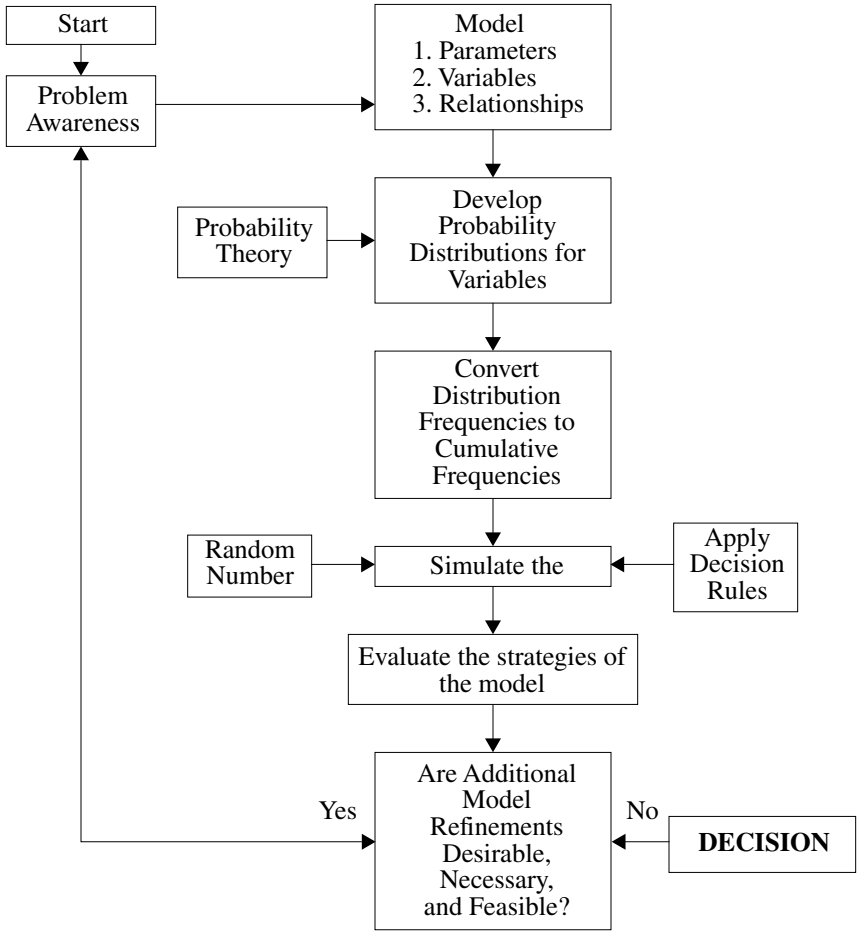


FIGURE H.1 Monte Carlo simulation.

numbers is used to describe the movement of each random variable over time. The random numbers allow an artificial but realistic sequence of events to occur. Monte Carlo simulation permits the experimenter to determine how varied policies or organizational conditions will be modified by the behavior of random or transient influences. A general approach to solving problems by Monte Carlo simulation is contained in Figure H.1.

Monte Carlo simulation establishes a stochastic model of a real situation and then performs sampling experiments on the model. This technique generates a vast amount of data that might otherwise take a very long time to obtain. Following the generation of data, computations can be made and a problem solution derived.

The major steps in Monte Carlo simulation are as follows:

1. Make sure that you know the probability distributions of certain key variables of the problem must be known distributions. They may be

standard distributions such as the Poisson, normal, or exponential, or they may be empirical distributions obtained from historical records.

2. Convert the frequency distributions to cumulative probability distributions. This assures that only one variable value will be associated with a given random number.
3. Sample at random from the cumulative probability distributions to determine specific variable values to use in the simulation. A way to sample is to use numbers from a table of random numbers. The random numbers are inserted in the cumulative probability distributions to obtain specific variable values for each observation. The sequence of assigned random numbers will imitate the pattern of variation expected to be encountered.
4. Simulate the operation under analysis for a large number of observations. The appropriate number of replications is determined in the same manner as the appropriate size of a sample in an actual experiment in the real world. The ordinary statistical tests of significance can be used. With computerized simulation the size of the sample can be increased without difficulty, and it is economical to run large samples with very small sampling errors.

Everything depends on the choice of frequency distributions. Unless there is some assurance they have been picked well, the entire simulation can be worthless. Distributions can be obtained from historical records or experimentation or chosen *a priori* on a quasi-subjective basis.

Random numbers are the life blood of the method and are numbers of equal long run frequency. They completely lack sequential predictability. The randomness of tabulated numbers can be validated by a chi-square test. The stream of random numbers can be obtained from a published table, or a computer can generate effectively random numbers (called pseudo-random numbers) internally.

Monte Carlo simulation has many practical uses, such as waiting line problems (where standard distributions for arrival rates and service rates are inadequate), layout problems of multiphase assembly lines, inventory problems, equipment replacement problems, engineering tolerancing and so on.

A simulation model does not produce an optimum solution. The experimenter selects the alternatives to evaluate by simulation but cannot be sure that the best alternative has been included. The simulation indicates possible solutions based only on the input of alternatives selected by the manager; it does not indicate which alternatives to evaluate. Simulation models usually develop heuristic rather than analytical solutions to a problem, but they can deal with very complex situations that defy solution by analytical methods.

No analytical solution can be extricated from its premises and assumptions. Simulation can investigate the effect of a relaxation of assumptions. Also, when no analytical solution is possible, simulation becomes important as a last resort. While simulation does not promise optimal solutions, it does allow picking out the best one tried. The ability of simulation to handle dependent variable interactions renders it a very powerful tool of systems analysis.

Simulation is used to reproduce a typical series of events (usually in mathematical form) that could have occurred in practice. If enough events are simulated and mean values determined, it can be assumed that they represent what would probably have happened in practice if the real situation existed. Standard statistical tests can be run on the output to determine when stability occurs.

Initial transient phenomena such as oscillations, rapid growth, and sudden decay are not unusual in simulation (or in reality). If system stability is desired, a sufficient startup period should be allowed for stability to develop. In real life, such transient phenomena are commonplace occurrences. Whereas analytical methods are usually based on steady-state conditions, simulation need not be limited by these assumptions.

SELECTED BIBLIOGRAPHY

Rubenstein, R.Y., *Monte Carlo Optimization, Simulation and Sensitivity of Queueing Networks*, John Wiley & Sons, New York, 1986.

Appendix I — Statistical Reporting Content

There are many ways to report the results of any statistical analysis. Here, we present a general outline that is functional and concise and clearly communicates the results.

Planning

- Clarify the objective.
- Develop a clear plan.
- Give yourself enough time.
- In this stage, think of the:
  - Executive summary
  - Problem description
  - Data description
  - Statistical methodology
  - Results and conclusions

Developing a report

- Write a quick first draft.
- Edit and proofread.
- Give your report a professional look.

Guidelines for effective reporting

Be clear.

- Provide sufficient background information.
- Tailor statistical explanations to your audience.
- Place charts and tables in the body of the report.

Be concise.

- Let charts do the talking.
- Be selective in the computer outputs you include.

Be precise.

- List assumptions and potential limitations.
- Limit the decimal places.
- Report the result fairly.
- Get advice from an expert.

EXAMPLE OF A TYPICAL STATISTICAL REPORT FORMAT

Header information
- Name of company
- Date
- To
- From
- Subject

Executive summary

Data set
- Software used
- Sample size
- Preliminary analysis

Analysis
- Objectives
- Statistical methodology
- Summary of measures
- Appropriate graphs

Results
- Regression analysis and forecasting (if appropriate)
- Conclusions and recommendations

Selected Bibliography

Albright, S.C., Winston, W.L., Zappe, C.J., and Kolesar, P., *Managerial Statistics*, Duxbury, Pacific Grove, CA, 2000.

Archetti, F. and Shoen, F., A survey on the global optimization problem: general theory and computational approaches, *Annals of Operation Research*, 1, 87–110, 1984.

Barnett, V., *Comparative Statistical Inference,* 3rd ed., Jossey-Bass, San Francisco, 2001.

Bather, J., *Decision Theory: An Introduction to Dynamic Programming and Sequential Decisions*, Jossey-Bass, San Francisco, 2001.

Bernardo, J.M., *Bayesian Theory*, Jossey-Bass, San Francisco, 2001.

Billingsley, P., *Probability and Measure*, Jossey-Bass, San Francisco, 2001.

Bloomfield, P., *Fourier Analysis of Time Series*, Jossey-Bass, San Francisco, 2001.

Chan, Saltelli and Scott, E.M., *Sensitivity Analysis*, Jossey-Bass, San Francisco, 2001.

Chao, X., Miyazawa, M., and Pinedo, M., *Queueing Networks: Customers, Signals and Product Form Solutions,* Jossey-Bass, San Francisco, 2001.

Chatterjee, S., Hadi, A.S., and Price, B, *Regression Analysis by Example*, Jossey-Bass, San Francisco, 2001.

Cohen, J. and Cohen, P., *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences,* Lawrence Erlbaum Associates, Hillsdale, NJ, 1983.

Cooper, R.B., *Introduction to Queuing Theory*, McMillan, New York, 1972.

Evans, M., Hastings, N., and Peacock, B., *Statistical Distributions,* 3rd ed., Jossey-Bass, San Francisco, 2001.

Finney, D.J., Standard errors of yields adjusted for regression on an independent measurement, *Biometrics Bulletin,* 2, 53–55, 1946.

Finney, J.M., Indirect effects in path analysis, *Sociological Methods and Research,* 1, 175–186, 1972.

Firebaugh, G., A rule for inferring individual-level relationships from aggregate data, *American Sociological Review,* 43, 557–572, 1978.

Firebaugh, G., Assessing group effects: A comparison of two methods, *Sociological Methods and Research,* 7, 384–395, 1979.

Fisher, F.M., *The Identification Problem in Econometrics*, McGraw-Hill, New York, 1966.

Fisher, R.A., The use of multiple measurements in taxonomic problems, *Annals of Eugenics,* 7, 179–198, 1936.

Fisher, R.A., *Statistical Methods for Research Workers*, 13th ed., Hafner, New York, 1958.

Fisher, R.A. and Yates, F., *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed., Hafner, New York, 1963.

Fletcher, R., *Practical Methods of Optimization,* 2nd ed., Jossey-Bass, San Francisco, 2001.

Forbes, D. and Tufte, E.R., A note of caution in causal modeling, *American Political Science Review,* 62, 1258–1264, 1968.

Fox, J., Effect analysis in structural equation models, *Sociological Methods and Research,* 9, 3–28, 1980.

Fox, K.A., *Intermediate Economic Statistics*, John Wiley & Sons, New York, 1968.

Freedman, J.L., Involvement, discrepancy, and change, *Journal of Abnormal and Social Psychology*, 69, 290–295, 1964.

Freund, R.J. and Little, R.C., *SAS System for Regression*, 3rd ed., Jossey-Bass, San Francisco, 2001.

Gaito, J., Repeated measurements designs and counterbalancing, *Psychological Bulletin,* 58, 46–54, 1961.
Games, P.A., Multiple comparisons of means, *American Educational Research Journal,* 8, 531–565, 1971.
Games, P.A., Limitations of analysis of covariance on intact group quasi-experimental designs, *Journal of Experimental Education,* 44, 51–54, 1976.
Gnanadesikan, R., *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley & Sons, New York, 1977.
Gocka, E.F., Stepwise regression for mixed mode predictor variables, *Educational and Psychological Measurement,* 33, 319–325, 1973.
Goldberger, A.S., *Econometric Theory,* John Wiley & Sons, New York, 1964.
Goldberger, A.S., On Boudon's method of linear causal analysis, *American Sociological Review,* 35, 97–101, 1970.
Goldberger, A.S. and Duncan, O.D., Eds., *Structural Equation Models in the Social Sciences,* Seminar Press, New York, 1973.
Gordon, R.A., Issues in multiple regression, *American Journal of Sociology,* 73, 592–616, 1968.
Gorsuch, R.L., *Factor Analysis*, W.B. Saunders, Philadelphia, 1974.
Graybill, F.A., *An Introduction to Linear Statistical Models*, Vol. 3, McGraw-Hill, New York, 1961.
Green, P.E., *Mathematical Tools for Applied Multivariate Analysis*, Academic Press, New York, 1976.
Green, P.E., *Analyzing Multivariate Data*, Dryden Press, Hinsdale, IL, 1978.
Green, P.E., Halbert, M.H., and Robinson, P.J., Canonical analysis: an exposition and illustrative application, *Journal of Marketing Research,* 3, 32–39, 1966.
Greene, V.L., An algorithm for total and indirect causal effects, *Political Methodology,* 4, 369–381, 1977.
Greenhouse, S.W. and Geisser, S., On methods in the analysis of profile data, *Psychometrika,* 24, 95–112, 1959.
Greenwald, A.G., Within-subjects designs: to use or not to use? *Psychological Bulletin,* 83, 314–320, 1976.
Gross, D. and Harris, C.M., *Fundamentals of Queueing Theory,* 3rd ed., Jossey-Bass, San Francisco, 2001.
Guilford, J.P., *Psychometric Methods,* 2nd ed., McGraw-Hill, New York, 1954.
Guilford, J.P. and Fruchter, B., *Fundamental Statistics in Psychology and Education,* 5th ed., McGraw-Hill, New York, 1978.
Gunst, R.F. and Mason, R.L., Advantages of examining multicollinearities in regression analysis, *Biometrics,* 33, 249–260, 1977.
Hoaglin, D.C., Mosteller, F., and Tukey, J.W., Eds., *Understanding Robust and Exploratory Data Analysis*, Jossey-Bass, San Francisco, 2001.
Hocking, R.R., Misspecification in regression, *American Statistician,* 28, 39–40, 1974.
Hocking, R.R., The analysis and selection of variables in linear regression, *Biometrics,* 32, 1–49, 1976.
Hodges, S.D. and Moore, P.G., Data uncertainties and least squares regression, *Applied Statistics,* 21, 185–195, 1972.
Hohn, F.E., *Elementary Matrix Algebra*, 2nd ed., MacMillan, New York, 1964.
Hollandar, M. and Wolfe, D.A., *Nonparametric Statistical Methods,* 2nd ed., Jossey-Bass, San Francisco, 2001.
Holzinger, K.J. and Freeman, F.N., The interpretation of Burt's regression equation, *Journal of Educational Psychology,* 16, 577–582, 1925.

Hooker, J., *Logic-Based Methods for Optimization: Combining Optimization and Constraint Satisfaction*, Jossey-Bass, San Francisco, 2001.

Hovel, A.E. and Kennard, R.W., Ridge regression: applications to nonorthogonal problems, *Technometrics,* 12, 69–82, 1970.

Horel, A.E. and Kennard, R.W., Ridge regression: biased estimation for nonorthogonal problems, *Technometrics,* 12, 55–67, 1970.

Hornbeck, F.W., Factorial analyses of variance with appended control groups, *Behavioral Science,* 18, 213–220, 1973.

Horst, P., *Matrix Algebra for Social Scientists*, Holt, Rinehart & Winston, New York, 1963.

Horst, P., *Psychological Measurement and Prediction,* Wadsworth, Belmont, CA, 1966.

Hottelling, H., Relations between two sets of variates, *Biometrika,* 28, 321–377, 1936.

Howe, H., Education research: the promise and the problem, *Educational Researcher,* 5(6), 2–7, 1976.

Huba, G.J., Wingard, J.A., and Bentler, P.M., A comparison of two latent variable causal models for adolescent drug use, *Journal of Personality and Social Psychology,* 40, 180–193, 1981.

Huberty, C.J., Multivariate indices of strength of association, *Multivariate Behavioral Research,* 7, 523–526, 1972.

Huberty, C.J., Discriminant analysis, *Review of Educational Research,* 45, 543–598, 1975.

Huberty, C.J., The stability of three indices of relative variable contribution in discriminant analysis, *Journal of Experimental Education,* 44, 59–64, 1975.

Huitema, B.E., *The Analysis of Covariance and Alternatives*, John Wiley and Sons, New York, 1980.

Hull, C.H. and Nie, N.H., *SPSS Update,* McGraw-Hill, New York, 1979.

Hummel, T.J. and Sligo, J.R., Empirical comparison of univariate and multivariate analysis of variance procedures, *Psychological Bulletin,* 76, 49–57, 1971.

Humphreys, L.G., Doing research the hard way: substituting analysis of variance for a problem in correlational analysis, *Journal of Educational Psychology,* 70, 873–876, 1978.

Humphreys, L.G. and Fleishman, A., Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables, *Journal of Educational Psychology,* 66, 464–472, 1974.

Huynh, H. and Feldt, L.S., Conditions under which mean square ratios in repeated measurements designs have exact F-distributions, *Journal of the American Statistical Association,* 65, 1582–1589, 1970.

Huynh, H. and Feldt, L.S., Estimation of the Box correction for degrees of freedom from sample data in randomized blocks and split-plot designs, *Journal of Educational Statistics,* 1, 69–82, 1976.

Johnston, J., *Econometric Methods,* 2nd ed., McGraw-Hill, New York, 1972.

Kemeny, J.G., Snell, J.L., and Thompson, G.L., *Introduction to Finite Mathematics,* 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1966.

Kenny, D.A., *Correlation and Causality,* John Wiley & Sons, New York, 1979.

Kerlinger, F.N., *Foundations of Behavioral Research,* 2nd ed., Holt, Rinehart & Winston, New York, 1973.

Khattree, R. and Naik, D., *Multivariate Data Reduction and Discrimination with SAS Software,* Jossey-Bass, San Francisco, 2001.

Kim, J.O. and Mueller, C.W., Standardized and unstandardized coefficients in causal analysis, *Sociological Methods and Research,* 4, 423–438, 1976.

Kirk, R.E., *Experimental Design: Procedures for the Behavioral Sciences*, Brooks/Cole, Belmont, CA, 1968.

Klecka, W.R., *Discriminant Analysis,* Sage, Beverly Hills, CA, 1980.

Kmenta, J., *Elements of Econometrics,* MacMillan, New York, 1971.
Krus, D.J., Reynolds, T.J., and Krus, P.H., Rotation in canonical variate analysis, *Educational and Psychological Measurement,* 36, 725–730, 1976.
Kumer, T.K., Multicollinearity in regression analysis, *Review of Economics and Statistics,* 57, 365–366, 1975.
Lana, R.E. and Lubin, A., The effect of correlation on the repeated measures design, *Educational and Psychological Measurement,* 23, 729–739, 1963.
Larsen, W.A. and McCleary, S.J., The use of partial residual plots in regression analysis, *Technometrics,* 14, 781–790, 1972.
Lerner, D., Ed., *Cause and Effect,* Free Press, New York, 1965.
Levine, M.S., *Canonical Analysis and Factor Comparison*, Sage, Beverly Hills, CA, 1977.
Li, C.C., *Introduction to Experimental Statistics*, McGraw-Hill, New York, 1964.
Li, C.C., *Path Analysis: A Primer,* Boxwood Press, Pacific Grove, CA, 1975.
Li, J.C.R., *Statistical Inference*, Edwards Brothers, Ann Arbor, MI, 1964.
Lindeman, R.H., Merenda, P.F., and Gold, R.Z., *Introduction to Bivariate and Multivariate Analysis,* Scott Foresman, Glenview, IL, 1980.
Linn, R.L., Fair test use in selection, *Review of Educational Research,* 43, 139–161, 1973.
Linn, R.L. and Werts, C.E., Errors of inference due to errors of measurement, *Educational and Psychological Measurement,* 33, 531–543, 1973.
Long, J.S., Estimation and hypothesis testing in linear models containing measurement error, *Sociological Methods and Research,* 5, 157–206, 1976.
Lord, F.M., Significance test for a partial correlation corrected for attenuation, *Educational and Psychological Measurement,* 34, 211–220, 1974.
MacDonald, K.I., Interpretation of residual paths and decomposition of variance, *Sociological Methods and Research,* 7, 289–304, 1979.
McLaclan, G. and Peel, D., *Finite Mixture Models*, Jossey-Bass, San Francisco, 2001.
Miller, R., *Optimization: Foundations and Applications,* Jossey-Bass, San Francisco, 2001.
Montgomery, D., Peck, E.A., and Vining, G., *Introduction to Linear Regression Analysis,* 3rd ed., Jossey-Bass, San Francisco, 2001.
Morrison, D.E. and Henkel, R.E., Eds., *The Significance Test Controversy: A Reader,* Aldine, Chicago, 1970.
Morrison, D.F., *Multivariate Statistical Methods,* 2nd ed., McGraw-Hill, New York, 1976.
Mulaik, S.A., *The Foundations of Factor Analysis,* McGraw-Hill, New York, 1972.
Myers, J.L., *Fundamentals of Experimental Design,* 3rd ed., Allyn & Bacon, Boston, 1979.
Myers, R.H. and Montomery, D.C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, Jossey-Bass, San Francisco, 2001.
Namboodiri, N.K., Ed., *Survey Sampling and Measurement,* Academic Press, New York, 1978.
Namboodiri, N.K., Carter, L.F., and Blalock, H.M., *Applied Multivariate Analysis and Experimental Designs,* McGraw-Hill, New York, 1975.
Noble, B. and Daniel, J.W., *Applied Linear Algebra*, 2nd ed., Prentice Hall, Englewood Cliffs, NJ, 1977.
Olson, C.L., On choosing a test statistic in multivariate analysis of variance, *Psychological Bulletin,* 83, 579–586, 1976.
Overall, J.E. and Klett, C.J., *Applied Multivariate Analysis,* McGraw-Hill, New York, 1972.
Overall, J. E., Spiegel, D.K., and Cohen, J., Equivalence of orthogonal and nonorthogonal analysis of variance, *Psychological Bulletin,* 82, 182–186, 1975.
Overall, J. E. and Woodward, J.A., Common misconceptions concerning the analysis of covariance, *Multivariate Behavioral Research,* 12, 171–186, 1977.
Pedhazur, E.J., Coding subjects in repeated measures designs, *Psychological Bulletin,* 84, 298–305, 1977.

Pillai, K.C. S., *Statistical Tables for Tests of Multivariate Hypotheses,* University of the Philippines, Manila, 1960.

Pincus, M., A Monte Carlo method for the approximate solution of certain types of constraint optimization problems, *Operations Research*, 18, 1225–1228, 1970.

Poor, D.D.S., Analysis of variance for repeated measures designs: two approaches, *Psychological Bulletin,* 80, 204–209, 1973.

Press, S.J., *Applied Multivariate Analysis*, Holt, Rinehart & Winston, New York, 1972.

Price, B., Ridge regression: application to nonexperimental data, *Psychological Bulletin,* 84, 759–766, 1977.

Raiffa, H. and Schlaifer, R., *Applied Statistical Decision Theory*, Jossey-Bass, San Francisco, 2001.

Reichardt, C.S., The statistical analysis of data from nonequivalent group designs, in *Quasi-experimentation,* Cook, T.D. and Campbell, D.T., Eds., Rand McNally, Skokie, IL, 1979.

Rockwell, R.C., Assessment of multicollinearity, *Sociological Methods and Research,* 3, 308–320, 1975.

Rogosa, D., Comparing nonparallel regression lines, *Psychological Bulletin,* 88, 307-321, 1980.

Rogosa, D., On the relationship between the Johnson-Neyman region of significance and statistical tests of parallel within-group regressions, *Educational and Psychological Measurement,* 41, 73–84, 1981.

Rohatgi, V.K. and Saleh, E., *An Introduction to Probability and Statistics,* 2nd ed., Jossey-Bass, San Francisco, 2001.

Rouanet, H. and Lepine, D., Comparison between treatments in a repeated-measures design: ANOVA and multivariate methods, *British Journal of Mathematical and Statistical Psychology,* 23, 147–163, 1970.

Roy, S.N., *Some Aspects of Multivariate Analysis*, John Wiley & Sons, New York, 1957.

Rozeboom, W.W., The fallacy of the null-hypothesis significance test, *Psychological Bulletin,* 57, 416–428, 1960.

Rozeboom, W.W., Ridge regression: bonanza or beguilement? *Psychological Bulletin,* 86, 242–249, 1979.

Rubin, D.B., Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology,* 66, 688–701, 1974.

Rubin, D.B., Assignment to treatment group on the basis of a covariate, *Journal of Educational Statistics,* 2, 1–26, 1977.

Rubinstein, R.Y., *Simulation and the Monte Carlo Methods,* John Wiley & Sons, New York, 1981.

Rubinstein, R.Y., *Monte Carlo Optimization, Simulation and Sensitivity of Queueing Networks*, John Wiley & Sons, New York, 1986.

Rulon, P.J. and Brooks, W.D., On statistical tests of group differences, in *Handbook of Measurement and Assessment in Behavioral Sciences,* Whitla, D.K., Ed., Addison-Wesley, Reading, MA, 1968.

Rulon, P.J., Tiedeman, D.V., Tatsuoka, M.M., and Langmuir, C.R., *Multivariate Statistics for Personnel Classification,* John Wiley & Sons, New York, 1967.

Ryan, T.A., Comments on orthogonal components, *Psychological Bulletin,* 56, 394-396, 1959.

Scheffe, H., *The Analysis of Variance,* John Wiley & Sons, New York, 1959.

Scheffler, I., Explanation, prediction, and abstraction, *British Journal for the Philosophy of Science,* 7, 293–309, 1957.

Schimek, M.G., Ed., *Smoothing and Regression: Approaches, Computation and Application*, Jossey-Bass, San Francisco, 2001.

Searle, S.R. and McCuloch, C.E., *Generalized, Linear and Mixed Models,* Jossey-Bass, San Francisco, 2001.
Specht, D.A., On the evaluation of causal models, *Social Science Research,* 4, 113–133, 1975.
Specht, D.A. and Warren, R.D., Comparing causal models, in *Sociological Methodology 1976,* Heise, D.R., Ed., Jossey-Bass, San Francisco, 1975.
Stevens, J.P., Four methods of analyzing between variation for the k-group MANOVA problem, *Multivariate Behavioral Research,* 7, 499–522, 1972.
Stevens, J.P., Global measures of association in multivariate analysis of variance, *Multivariate Behavioral Research,* 7, 373-378, 1972.
Stimson, J.A., Carmines, E.G., and Zeller, R.A., Interpreting polynomial regression, *Sociological Methods and Research,* 6, 515-524, 1978.
Stolzenberg, R.M., The measurement and decomposition of causal effects in nonlinear and nonadditive models, in *Sociological Methodology 1980,* Schuessler, K.F., Ed., Jossey-Bass, San Francisco, 1979.
Sudman, S., *Applied Sampling,* Academic Press, New York, 1976.
Sullivan, J.L. and Feldman, S., *Multiple Indicators: An Introduction,* Sage, Beverly Hills, CA, 1979.
Tatsuoka, M.M., *Discriminant Analysis,* Institute for Personality and Ability Testing, Champaign, IL, 1970.
Tatsuoka, M.M., *Significance Tests: Univariate and Multivariate,* Institute for Personality and Ability Testing, Champaign, IL, 1971.
Tatsuoka, M.M., *Classification Procedures: Profile Similarity,* Institute for Personality and Ability Testing, Champaign, IL, 1974.
Tatsuoka, M.M., Classification procedures, in *Introductory Multivariate Analysis,* Amick, D.J. and Walberg, H.J, Eds., McCutchan, Berkeley, CA, 1975.
Tatsuoka, M.M., Discriminant analysis., in *Data Analysis Strategies and Designs for Substance Abuse Research,* Bentler, P.M., Lettieri, D.L, and Austin, G.A., Eds., U.S. Government Printing Office, Washington, DC, 1976.
Taylor, C.L., Ed., *Aggregate Data Analysis,* Mouton, Paris, 1968.
Thorndike, R.M., *Correlational Procedures for Research*, Gardner Press, New York, 1978.
Thorndike, R.M. and Weiss, D.J., A study of the stability of canonical correlations and canonical components, *Educational and Psychological Measurement,* 33, 123–134, 1973.
Timm, N.H., *Multivariate Analysis with Applications in Education and Psychology,* Brooks/Cole, Monterey, CA, 1975.
Tucker, R.K. and Chase, L.J., Canonical correlation, in *Multivariate Techniques in Human Communication Research,* Monge, P.R. and Cappella, J.N., Eds., Academic Press, New York, 1980.
Tukey, J.W., Analyzing data: sanctification or detective work? *American Psychologist,* 24, 83–91, 1969.
Van Ryzin, J., Ed., *Classification and Clustering,* Academic Press, New York, 1977.
Velicer, W.F., Suppressor variables and the semipartial correlation coefficient, *Educational and Psychological Measurement,* 38, 953–958, 1978.
Walker, H.M. and Lev, J., *Statistical Inference,* Holt, Rinehart & Winston, New York, 1953.
Warren, W.G., Correlation or regression: bias or precision, *Applied Statistics,* 20, 148–164, 1971.
Winer, B.J., *Statistical Principles in Experimental Design*, 2nd ed., McGraw-Hill, New York, 1971.
Wu, C.F. and Hamada, M., *Experiments: Planning, Analysis, and Parameter Design Optimization*, John Wiley & Sons, New York, 2000.

Index

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Features

- Provides a plethora of statistical techniques, tests, and methodologies used in the Six Sigma Methodology
- Covers basic parametric and nonparametric statistics
- Reviews probability concepts and applications
- Introduces the relationship of statistics, probability, and reliability as they apply to quality and to Six Sigma Methodology
- Focuses on understanding and interpreting statistical results

Six Sigma and Beyond: Statistics and Probability, Volume III introduces the relationship of statistics, probability, and reliability as they apply to quality in general and to Six Sigma in particular. Researchers and professionals in all walks of life recognize the value of utilizing probability and statistics, but often do not have the necessary experience in both concept and application. This is also true in the field of quality and especially in six sigma methodology. This valuable reference covers the concepts of many useful statistical tools, appropriate formulae for these specific tools, the connection of statistics to probability, and how to use them.

The author brings the theoretical into the practical by providing statistical techniques, tests, and methods that the reader can use in any organization. He reviews basic parametric and nonparametric statistics, probability concepts and applications, and addresses topics for both measurable and attribute characteristics. He delineates the importance of collecting, analyzing, and interpreting data not from an academic point of view but from a practical perspective. **Six Sigma and Beyond** shows you how to use statistical tools to improve your processes and give your organization the competitive edge.

SL3127

ISBN 1-57444-312-7

90000

9 781574 443127

St. Lucie Press

www.crcpress.com