Cheng-Few Lee
John C. Lee
Alice C. Lee

Statistics for Business and Financial Economics

Third Edition

Springer

Statistics for Business and Financial Economics

Cheng-Few Lee • John C. Lee • Alice C. Lee*

Statistics for Business and Financial Economics

Third Edition

*Disclaimer: Any views or opinions presented in this publication are solely those of the authors and do not necessarily represent those of State Street Corporation. State Street Corporation is not associated in any way with this publication and accepts no liability for the contents of this publication.

Springer

Cheng-Few Lee
Department of Finance and Economics
Rutgers University Business School
Piscataway, New Jersey
USA

John C. Lee
Center for PBBEF Research
Morrisplains, New Jersey
USA

Alice C. Lee
State Street Corporation
Boston, Massachusetts
USA

ISBN 978-1-4614-5896-8 ISBN 978-1-4614-5897-5 (eBook)
DOI 10.1007/978-1-4614-5897-5
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012951347

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

We like to dedicate this book to
Schwinne C. Lee, Jennifer Lee, Michael Lee,
and Michelle Lee.

Cheng-Few Lee, John C. Lee,
and Alice C. Lee

About the Authors

Cheng-Few Lee is a Distinguished Professor of Finance at Rutgers Business School, Rutgers University and was chairperson of the Department of Finance from 1988–1995. He has also served on the faculty of the University of Illinois (IBE Professor of Finance) and the University of Georgia. He has maintained academic and consulting ties in Taiwan, Hong Kong, China and the United States for the past three decades. He has been a consultant to many prominent groups including, the American Insurance Group, the World Bank, the United Nations and The Marmon Group Inc., Wintek Corporation and Polaris Financial Group, etc.

Professor Lee founded the Review of Quantitative Finance and Accounting (RQFA) in 1990 and the Review of Pacific Basin Financial Markets and Policies (RPBFMP) in 1998, and serves as managing editor for both journals. He was also a co-editor of the Financial Review (1985–1991) and the Quarterly Review of Economics and Business (1987–1989).

In the past thirty-nine years, Dr. Lee has written numerous textbooks ranging in subject matter from financial management to corporate finance, security analysis and portfolio management to financial analysis, planning and forecasting, and business statistics. Dr. Lee has also published more than 200 articles in more than twenty different journals in finance, accounting, economics, statistics, and management. Professor Lee has been ranked the most published finance professor worldwide during 1953–2008.

Alice C. Lee is currently a vice president in finance at State Street Corporation, heading up a group that provides analytics and valuations in support to the corporate Chief Accounting Officer. She was also previously a Vice President in the Model Validation Group, Enterprise Risk Management, at State Street Corporation. Her career spans over 20 years of experience, with a diverse background that includes academia, engineering, sales, and management consulting. Her primary areas of expertise and research are corporate finance and financial institutions. She is coauthor of Statistics for Business and Financial Economics, 2nd ed and 3rd ed (with Cheng F. Lee and John C. Lee), Financial Analysis, Planning and Forecasting, 2nd ed (with Cheng F. Lee and John C. Lee), and Security Analysis, Portfolio Management, and Financial Derivatives (with Cheng F. Lee, Joseph Finnerty,

John C. Lee and Donald Wort). In addition, she has coedited other annual publications including Advances in Investment Analysis and Portfolio Management (with Cheng F. Lee).

John C. Lee is a Microsoft Certified Professional in Microsoft Visual Basic and Microsoft Excel VBA. He has a Bachelor and Masters degree in accounting from the University of Illinois at Urbana-Champaign. John has worked over 20 years in both the business and technical fields as an accountant, auditor, systems analyst and as a business software developer. He is the author of the book on how to use MINITAB and Microsoft Excel to do statistical analysis which is a companion text to Statistics of Business and Financial Economics, of which he is one of the co-authors. In addition, he also published Financial Analysis, Planning and Forecasting, 2ed. (with Cheng F. Lee and Alice C. Lee) , and Security Analysis, Portfolio Management, and Financial Derivatives (with Cheng F. Lee, Joseph Finnerty, Alice C. Lee and Donald Wort). John has been a Senior Technology Officer at the Chase Manhattan Bank, Assistant Vice President at Merrill Lynch and Associated Director at UBS. Currently, he is the Director of the Center for PBBEF Research.

Preface to the Third Edition

Since the first edition of this book was published in 1993, and the second edition was published in 2000, it has been widely used in universities in the United States, Asia, Europe, and other countries. The following universities had adopted this book as a course text (Here is a partial list of the schools that have adopted this statistics book. However, it is not a full list because publishers do not have access to the wholesaler's list of schools that purchase this book):

Aarhus School of Business, Denmark	State University of New York – Binghamton University, USA
University of Alabama, USA	Norwegian School of Economics & Business Administration, Norway
Aoyama Gakun University, Japan	University of North Carolina at Greensboro, USA
University of Arkansas, USA	University of Notre Dame, USA
Bogazici University, Turkey	Reading University, England, UK
University of California, Los Angeles, USA	Rutgers University, USA
Carnegie Mellon University, USA	San Francisco State University, USA
Chaminade University of Honolulu, USA	St Joseph's College-Suffolk Campus, USA
Catholic University of America, USA	University of St. Thomas, USA
National Cheng Kung University, Taiwan	Suffolk University, USA
Cleary University, USA	National Taiwan University, Taiwan
National Chiao Tung University, Taiwan	Virginia Polytechnic & State University, USA
University of Gothenburg, Sweden	Washington University, USA
City University of Hong Kong, China	Western Kentucky University, USA
University of Hartford, USA	Western Washington University, USA
University of Illinois at Chicago, USA	
University of Illinois Medical Center, USA	
Kainan University, Taiwan	
Northern Illinois University, USA	
Monmouth University, USA	
New York University, USA	

We appreciate the schools that use the Second Edition and who have given us useful suggestions to improve this book. To the best of our knowledge, this is the only business statistics book that uses finance, economic, and accounting data throughout the entire book. Therefore, this book gives students an understanding of how to apply the methodology of statistics to real-world situations. In particular, this book shows how descriptive statistics, probability, statistical distributions, statistical inference, regression methods, and statistical decision theory can be used to analyze individual stock price, stock index, stock rate of return, market rate of return, and decision making. In addition, this book also shows how time-series analysis and the statistical decision theory method can be used to analyze accounting and financial data.

How This Edition Has Been Revised

In this edition, we first update the real-world examples and revise some sections to improve the ease understanding the topics. The auto companies, GM and Ford, used in empirical section of each chapter are replaced by two pharmaceutical firms, Johnson & Johnson and Merck. We update the data of stock price, dividend per share, earnings per share, and financial ratios of Johnson & Johnson and Merck until 2010. The annual macroeconomic data, such as prime rate, GDP, CPI, 3-month T-Bill rate, are updated to 2009. The EPS, DPS, and PPS for Dow Jones 30 Industrial Firms used in the project are also updated to 2009. The time aggregation and the estimation of the market model are added in example 16.8. The questions added to this edition are as follows:

Chapter	Problems
1	28, 29, 30, 31
2	52, 53, 54, 55
3	50, 51, 52, 53
4	63, 64, 65, 66, 67, 68, 69, 70
5	83, 84, 85, 86
6	75, 76, 77, 78
7	70, 71, 72, 73
8	88, 89, 90, 91, 92
9	68, 69, 70, 71
10	102, 103, 104, 105
11	100, 101,102,103, 104
12	99, 100, 101, 102
13	77, 78, 79, 80, 81
14	70, 71, 72, 73, 74
15	66, 67, 68, 69, 70
16	72, 73, 74, 75, 76
17	82, 83, 84, 85, 86
18	77, 78, 79, 80, 81

(continued)

(continued)

Chapter	Problems
19	64, 65, 66, 67, 68
20	86, 87, 88, 89, 90
21	68, 69, 70, 71, 72, 73

Alternative Ways to Use the Text

There *are **five*** alternative approaches to use the new edition of this book. They can be described as follows:

A. *Traditional Approach*
The goal of this approach is to demonstrate to the students the basic applications of statistics in general business, economics, and finance. This goal can be achieved by skipping all appendices, technical footnotes, optional sections, and other sections at the instructor's discretion. Using this alternative, students need only basic algebra, geometry, and business and economic common sense to understand how statistics can be used in general business, economics, and finance applications.

B. *Accounting and Financial Data Analysis Approach*
The goal of this approach is not only to illustrate basic overall business, economic, and finance applications but to show how to use statistics in accounting and financial data analysis and decision making. This goal can be achieved by omitting all the technical appendices, technical footnotes, and most optional sections but covering all or most of the following topics:

Chapter	Topic
2	Appendices 2 and 3 on stock market rates of return and on financial statements and financial ratio analysis
4	Appendix 3, financial ratios for two pharmaceutical firms
6	Appendix 2, applications of the binomial distribution to evaluate call options
7	Appendix 2, cumulative normal distribution function and the option pricing model
9	Section 9.8, analyzing the first four moments of rates of return of the 30 DJI firms
10	Appendix 1, a control chart approach for cash management
13	Appendix 1, derivation of normal equations and optimal portfolio weights
13	Appendix 4, American call option and bivariate normal CDF
16	Appendix 1, dynamic ratio analysis; Appendix 2, term structure of interest rate
19	Section 19.5, stock market indexes; Appendix 1, options on stock indexes and currencies; Appendix 2, index futures and hedge ratio
21	Sections 21.7 and 21.8 on mean and variance trade-off analysis and the mean and variance method for capital budgeting decisions; Appendices 2, 3, and 4 on the graphical derivation of the standard deviation for NPV

C. *Project Approach*

Based upon all five projects, the instructor can use the project approach to teach the course. Under this approach, the instructor can ask students to write a term project by using accounting, economic, and financial data collected from Yahoo Finance and St. Louis Federal Reserve Bank. The five projects are as follows:

Project I: Project for descriptive statistics
Project II: Project for probability and important distributions
Project III: Project for statistical inferences based on samples
Project IV: Project for regression and correlation analyses
Project V: Project for selected topics in statistical analysis

D. *Calculus Approach*

The objective of the fourth approach is to show students how calculus can be used in statistical analysis. To achieve this goal, the instructor can try to cover all optional sections and as many of the technical footnotes and appendices as possible. To do this, of course, the instructor may have to skip many application examples, such as the finance applications discussed in Approach B.

E. *Financial Analysis, Planning and Forecasting Approach*

This book can be used for a course entitled Financial Analysis, Planning and Forecasting by covering every topic presented in Chapters 2, 3, 4, 6, 7, 9, 13, 14, 15, 16, 18, 19, and 21.

In addition to using this book as a textbook, it can also be very useful as a reference book for managers who deal with accounting and financial data analysis.

We would like to recommend that the instructor consider requiring students to solve the following problems by using either MINITAB, Microsoft Excel, or SAS programs:

Chapter	Problems
2	7, 23
3	22, 25, 30, 50, 51, 53
4	4, 6, 7, 8, 27, 38, 39, 40, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 64
6	8, 12, 73, 74
7	5, 43
8	7, 85, 86, 87
9	35, 39, 48
10	27, 28, 55, 104, 105
11	5, 9, 46, 98, 99
12	3, 20, 21, 22, 23, 44, 84, 99, 100, 101, 102
13	5, 10, 23, 47, 48, 49, 50, 51, 63, 64, 65, 66, 67, 68, 69, 70, 78, 79
14	7, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 40, 65, 70, 74
15	10, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 70
16	27, 28, 31, 34, 35, 38, 41, 42, 43, 44, 45, 66, 67, 68, 75, 76
17	17, 19, 39, 40, 41, 42, 63
18	7, 34, 35, 36, 37, 38, 39, 40, 41, 42, 50, 60, 61, 62, 64, 68, 69, 76, 77, 78, 79, 80, 81
19	62, 63, 68
20	72, 73
21	12, 65

Supplementary Materials

Study Guide, by Li-Shya Chen, National Chengchi University, Taiwan, Lie-Jane Kao, Kainan University, Taiwan, and Ronald L. Moy, St. John's University. This fine workbook encourages learning by doing. Each chapter begins with a section describing the basic concept of that chapter in intuitive terms. Then, the student goes on to a formal review of the chapter and several worked-out problems that show in details how the solution is obtained. A variety of multiple-choice, true-false, and open-ended questions and problems follows. All answers are included at the end of each chapter.

Data Sets. A wide variety of macroeconomic, financial, and accounting data is available on computer disks to facilitate student practice. A complete listing of these data sets is given at the end of this book. The disks themselves are free of charge.

Instructor's Guide. The three main parts of the Instructor's Guide are the Overview and Objectives; the complete solutions to the text problems by Cheng F. Lee, John C. Lee, Li-Shya Chen, Lie-Jane Kao; and the Test Bank, with more than 1,000 multiple-choice and true-false problems, by Alice C. Lee, Li-Shya Chen, and Lie-Jane Kao. Most instructors will find the Instructor's Guide indispensable.

Computerized Testing Program. With the Test Bank on CD-ROM for notebook or desktop computers, instructors can select, rearrange, edit, or add problems as they wish.

New Jersey, USA Cheng-Few Lee

Acknowledgments

For the third edition, we appreciate our secretaries, staff, and my assistants, including Ms. Miranda Mei-Lan Luo, Tzu Tai, Hong-Yi Chen, Anthony Gallo, for being very helpful in updating and typing the text for the new edition and the instructor's manual for this book. Finally, we like to thank the Wintek Corporation and APEX International Financial Engineering Res. & Tech. Co., Ltd for the financial support that allowed us to write this book.

Cheng F. Lee
John C. Lee
Alice C. Lee

Preface to the Second Edition

Since the first edition of this book was published in 1993, it has been widely used in universities in the United States, Asia, and Europe. The following universities had adopted this book as a course text:

Aarhus School of Business
University of Alabama
Aoyama Gakun University
University of Arkansas
University of California, Los Angeles
Carnegie Mellon University
Catholic University of America
National Cheng Kung University
City University of Hong Kong
University of Hartford
University of Illinois Medical Center
Northern Illinois University
Monmouth University
New York University
Norwegian School of Economics & Business Administration
University of North Carolina at Greensboro
University of Notre Dame
Reading University
Rutgers University
San Francisco State University
University of St. Thomas
Suffolk University
National Taiwan University
Virginia Polytechnic & State University
Washington University
Western Kentucky University
Western Washington University

How This Edition Has Been Revised

In addition to correction of errors, the new edition uses the most updated real-world data on accounting, finance, and economics. The most recent version of MINITAB (Version 12) has been used for most of the empirical examples. In addition, Microsoft Excel 97 has been explicitly introduced in this book. The new material added to this edition is briefly described as follows:

Appendix 2A	Microsoft Excel to Draw Graphs
Appendix 2B	Stock Rates of Return and Market Rates of Return
Appendix 2C	Financial Statements and Financial Ratio Analysis
Appendix 3A	Financial Ratio Analysis
Appendix 4C	Financial Ratio Analysis for Three Auto Firms
Appendix 7A	Mean and Variance for Continuous Random Variables
Appendix 7B	Cumulative Normal Distribution Function and the Option Pricing Model
Appendix 7C	Lognormal Distribution Approach to Derive the Option Pricing Model
Section 9.4	The Chi-Square Distribution and the Distribution of Sample Variance
Section 9.8	Analyzing the First Four Moments of Rates of Return of the 30 DJI Firms
Appendix 9E	Noncentral χ^2 and Option Pricing Model
Section 10.9	Control Charts for Quality Control
Section 11.3	Hypothesis Test Construction and Testing Procedure
Appendix 11A	The Power of a Test, the Power Function, and the Operating-Characteristic Curve
Appendix 12A	ANOVA and Statistical Quality Control
Appendix 13D	American Call Option and Bivariate Normal CDF
Appendix 16A	Dynamic Ratio Analysis
Appendix 16B	Term Structure of Interest Rate
Application 19.3	CPI, Inflation Rate, and Interest Rate
Appendix 19A	Options on Stock Indexes and Currencies
Appendix 19B	Index Futures and Hedge Ratio
Section 21.7	Mean and Variance Trade-Off Analysis
Appendix E	Useful Formula in Statistics
Appendix F	Important Finance Topics

In addition, a real-world application project is added to the end of each part to show how the topics discussed can be applied in analyzing the real-world financial data. They are:

Project I: Project for Descriptive Statistics
Project II: Project for Probability and Important Distributions
Project III: Project for Statistical Inferences Based on Samples
Project IV: Project for Regression and Correlation Analyses
Project V: Project for Selected Topics in Statistical Analysis

Alternative Ways to Use the Text

There *are five* alternative approaches to use the new edition of this book. They can be described as follows:

A. *Traditional Approach*
The goal of this approach is to demonstrate to the students the basic applications of statistics in general business, economics, and finance. This goal can be achieved by skipping all appendices, technical footnotes, optional sections, and other sections at the instructor's discretion. Using this alternative, students need only basic algebra, geometry, and business and economic common sense to understand how statistics can be used in general business, economics, and finance applications.

B. *Accounting and Financial Data Analysis Approach*
The goal of this approach is not only to illustrate basic overall business, economic, and finance applications but to show how to use statistics in accounting and financial data analysis and decision making. This goal can be achieved by omitting all the technical appendices, technical footnotes, and most optional sections but covering all or most of the following topics:

Chapter	Topic
2	Appendices 2 and 3 of Chap. 2 on stock market rates of return and on financial statements and financial ratio analysis
3	Appendix 1 of Chap. 3, financial ratio analysis
4	Appendix 3 of Chap. 4, financial ratios for three auto firms
6	Appendix 2 of Chap. 6, applications of the binomial distribution to evaluate call options
7	Appendix 2 of Chap. 7, cumulative normal distribution function and the option pricing model
9	Section 9.8, analyzing the first four moments of rates of return of the 30 DJI firms
10	Appendix 1 of Chap. 10, a control chart approach for cash management
Appendix 1 of Chap. 13	Derivation of normal equations and optimal portfolio weights
Appendix 4 of Chap. 13	American call option and bivariate normal CDF
16	Appendix 1 of Chap. 16, dynamic ratio analysis and Appendix 2 of Chap. 16, term structure of interest rate
19	Section 19.5, stock market indexes and Appendix 1 of Chap. 19, options on stock indexes and currencies. Appendix 2 of Chap. 19, index futures and hedge ratio
21	Sections 21.7 and 21.8 on mean and variance trade-off analysis and the mean and variance method for capital budgeting decisions; Appendices 2, 3, and 4 of Chap. 21 on the graphical derivation of the standard deviation for NPV

C. *Project Approach*
Based upon all five projects, the instructor can use the project approach to teach the course. Under this approach, the instructor can ask students to write a term project by using accounting, economic, and financial data.

D. *Calculus Approach*
The objective of the fourth approach is to show students how calculus can be used in statistical analysis. To achieve this goal, the instructor can try to cover all optional sections and as many of the technical footnotes and appendices as possible. To do this, of course, the instructor may have to skip many application examples, such as the finance applications discussed in Approach B.

E. *Financial Analysis, Planning and Forecasting Approach*
This book can be used for a course entitled *Financial Analysis*, *Planning and Forecasting* by covering every topic presented in Chaps. 2, 3, 4, 6, 7, 9, 13, 14, 15, 16, 18, 19 and 21.

In addition to using this book as a textbook, it can also be very useful as a reference book for managers who deal with accounting and financial data analysis.

We would like to recommend that the instructor consider requiring students to solve the following problems by using either MINITAB, Microsoft Excel, or SAS programs:

Chapter	Problems
2	7, 23
3	22, 25, 30
4	4, 6, 7, 8, 27, 38, 39, 40, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54
6	8, 12, 73, 74
7	5, 43
8	7, 85, 86, 87
9	35, 39, 48
10	27, 28, 55
11	5, 9, 46, 98, 99
12	3, 20, 21, 22, 23, 44, 84
13	5, 10, 23, 47, 48, 49, 50, 51, 63, 64, 65, 66, 67, 68, 69, 70
14	7, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 40, 65
15	10, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37
16	27, 28, 31, 34, 35, 38, 41, 42, 43, 44, 45, 66, 67, 68
17	17, 19, 39, 40, 41, 42, 63
18	7, 34, 35, 36, 37, 38, 39, 40, 41, 42, 50, 60, 61, 62, 64, 68, 69, 76
19	62, 63
20	72, 73
21	12, 65

Supplementary Materials

Study Guide, by Ronald L. Moy, St. John's University. This fine workbook encourages learning by doing. Each chapter begins with a section describing the basic concept of that chapter in intuitive terms. Then, the student goes on to a formal review of the chapter and several worked-out problems that show in details

how the solution is obtained. A variety of multiple-choice, true-false, and open-ended questions and problems follows. All answers are included at the end of each chapter.

MINITAB and Microsoft Excel Book, by John C. Lee, Chase Manhattan Bank. The book, which follows the textbook chapter by chapter, is designed to help students use MINITAB and (or) Microsoft Excel throughout the course. Each chapter includes a variety of specific applications and ends with a statistical summary.

Data Sets. A wide variety of macroeconomic, financial, and accounting data is available on computer disks to facilitate student practice. A complete listing of these data sets is given at the end of this book. The disks themselves are free of charge.

Instructor's Guide. The three main parts of the Instructor's Guide are the Overview and Objectives; the complete solutions to the text problems by Cheng F. Lee, John C. Lee, and Edward Bubnys; and the Test Bank, with more than 1,000 multiple-choice and true-false problems, by Alice C. Lee, Pricewaterhouse Coopers. Most instructors will find the Instructor's Guide indispensable.

Computerized Testing Program. With the Test Bank on disk for either IBM or Macintosh computers, instructors can select, rearrange, edit, or add problems as they wish.

New Jersey, USA Cheng-Few Lee
New Jersey, USA John C. Lee
Massachusetts, USA Alice C. Lee

Acknowledgments

For this new edition, suggestions from Professors Richard T. Baillie, Abdul Basti, Y. C. Chang, Dongcheol Kim, Ron Moy, Terry G. Seaks, Ed Bubnys, Chin-Chen Chien, and others are most appreciated. In addition, Ta-Peng Wu and Chingfu Chang's help is also appreciated.

My secretarial staff, including Gerry Leo, Bertha Martinez, and Nikki Lewis, have been very helpful in typing the text for the new edition and the instructor's manual for this book.

Cheng F. Lee
John C. Lee
Alice C. Lee

Preface to the First Edition

When I first began writing *Statistics for Business and Financial Economics*, my goal was to develop a text that would give my students at the University of Illinois and at Rutgers University the basic statistical tools they need not only for a general business school education but also for the statistics that a finance major needs. Over time, that original purpose has evolved into a broad statistical approach that integrates concepts, methods, and applications. The scope has widened to include all students of business and economics, especially upper-level undergraduates and MBA students, who want a clear and comprehensive introduction to statistics. This book is written for them.

A distinguishing feature of the text is the creative ways in which it weaves useful and interesting concepts from general business (accounting, marketing, management, and quality control), economics, and finance into the text. It actively shows how various statistical methods can be applied in business and financial economics.

More specifically, the text incorporates the following pedagogical features:

Usefulness of statistical methods. This text features an unusually large number of real-life examples that show students how statistical methods can help them.

Non-calculus approach. Extensive use of examples and applications (more than 500) in the text and problem sets at the end of the chapters (more than 1,500) shows students how statistical methodology can be effectively implemented and applied. All the examples, applications, and problems can be worked out using only high-school algebra and geometry. Calculus, which offers an alternative and intellectually satisfying perspective, is presented only in footnotes and appendixes.

Emphasis on data analysis. Most statistics texts, in their justifiable need to demonstrate to students how to use the various statistical tests, focus all too often on the mechanical aspects of problem solving. Lost is the simple but important notion that statistics is the study of data. Data analysis is an important theme of this text. In particular, one set of financial economic data for GM and Ford is used continuously throughout the text for various types of statistical analysis.

Use of computers. After students understand the step-by-step processes, the text shows how computers can make statistical analysis more efficient and less time consuming. Examples utilizing MINITAB, Lotus 1-2-3, and SAS are shown, and a supplementary manual based entirely on MINITAB is available.

Straightforward language. Not least, the text employs clear and simple language to guide the reader to a knowledge of the basic statistical methods used in business decision making and financial economics.

Additionally, this text explores in slightly greater depth many of the standard statistical topics: There is more coverage of regression analysis than in other texts (see Chaps. 13, 14, 15, and 16 and part of Chaps. 18, 19, 20, and 21). Quality control is explicitly integrated with point and interval estimation (Chap. 10). Stock market indexes and the index of leading economic indicators are both treated as an expanded portion of regular index numbers (Chap. 19).

Many chapters have appendixes that develop useful financial applications of the standard topics found in the chapter body. Some appendixes may be used as case studies and the following will especially serve the purpose:

Financial Statements and Financial Ratio Analysis (Appendix 3 of Chap. 2, Appendix 1 of Chap. 3, and Appendix 3 of Chap. 4 may be used together as a single case study)
Applications of the Binomial Distribution to Evaluate Call Options (Appendix 2 of Chap. 6)
Cumulative Normal Distribution Function and the Option Pricing Model (Appendix 2 of Chap. 7)
Control Chart Approach for Cash Management (Appendix 1 of Chap. 10)

Organization of the Text

The text has 21 chapters divided into five parts. Part I, Introduction and Descriptive Statistics, consists of four chapters. Following the introductory chapter, Chap. 2 addresses Data Collection and Presentation. Chapter 3 delves into Frequency Distributions and Data Analyses. It is followed by Numerical Summary Measures in Chap. 4.

Probability and Important Distributions, Part II, includes five chapters, the first of which, Chap. 5, is entitled Probability Concepts and Their Analysis. Discrete Random Variables and Probability Distributions are discussed in Chap. 6, after which Chap. 7 covers The Normal and Lognormal Distributions. Sampling and Sampling Distributions are covered in Chap. 8. Chapter 9 closes Part II of the text by discussing Other Continuous Distributions and Moments for Distributions.

Part III, Statistical Inferences Based on Samples, comprises three chapters. Chapter 10 covers Estimation and Statistical Quality Control. Chapter 11 explores Hypothesis Testing and Chap. 12 provides an Analysis of Variance and Chi-Square Tests.

Chapters 13, 14, 15, and 16 make up Part IV, which is entitled Regression and Correlation: Relating Two or More Variables. The first of these chapters is Simple Linear Regression and the Correlation Coefficient. From a discussion of Simple Linear Regression and Correlation: Analyses and Applications in Chap. 14, this book moves on to address Multiple Linear Regression in Chap. 15. Finally, Chap. 16 closes Part IV with a look at Other Topics in Applied Regression Analysis.

The last part of the text, Part V, considers Selected Topics in Statistical Analysis for Business and Economics. Nonparametric Statistics is the subject of Chap. 17, which is followed by an exploration of Time Series: Analysis, Model, and Forecasting in Chap. 18. Chapters 19 and 20 discuss Index Numbers and Stock Market Indexes, and Sampling Surveys: Methods and Applications, respectively. Statistical Decision Theory: Methods and Applications is the topic of the final chapter, Chap. 21.

There are four appendixes. Appendix A provides 14 statistical tables. Appendix 1 gives a full description of the data sets available on a computer disk. Appendix 2 briefly describes the use of MINITAB, especially the microcomputer version, and Appendix 3 introduces the microcomputer version of SAS. Finally, to make sure they are on the right track in working the problems, students can consult the section at the end of this book that gives short Answers to Selected Odd-Numbered Questions and Problems. (Full solutions are given in the Instructor's Guide.)

About This First Edition

One legitimate concern with a new statistics text is that the first edition will contain errors (too many errors!) that must await correction only in the second edition. We have taken action to confront this problem by carrying out a thorough and detailed accuracy check of the entire text: Every problem in the text has been reworked by "outsiders" to the project. So confident are we that this is an error-free book that the publisher is willing to pay $10 for the first report (in writing) of each substantive error.

Alternative Ways to Use the Text

Based upon my own teaching experience, I would like to suggest three alternative ways to use this textbook.

Alternative One: The goal of this alternative is to demonstrate to students the basic applications of statistics in general business, economics, and finance. This goal can be achieved by skipping all appendixes, technical footnotes, optional sections, and other sections at the instructor's discretion. Using this alternative, the student needs only

basic algebra, geometry, and business and economic common sense to understand how statistics can be used in general business, economics, and finance applications.

Alternative Two: The goal of this alternative is not only to illustrate basic overall business, economic, and finance applications but to show how to use statistics in financial analysis and decision making. This goal can be achieved by omitting all the technical appendixes, technical footnotes, and most optional sections but covering all or most of the following topics:

Chapter	Topic
2	Appendices 2 and 3 of Chap. 2 on stock and market rates of return and on financial statements and financial ratio analysis
3	Appendix 1 of Chap. 3, financial ratio analysis
4	Appendix 3 of Chap. 4, financial ratios for three auto firms. As mentioned earlier, Appendix 3 of Chap. 2, Appendix 1 of Chap. 3, and Appendix 3 of Chap. 4 can be treated as a single case study
6	Appendix 2 of Chap. 6, applications of the binomial distribution to evaluate call options
7	Appendix 2 of Chap. 7, cumulative normal distribution function and the option pricing model
9	Section 9.8, analyzing the first four moments of rates of return of the 30 DJI firms
10	Appendix 1 of Chap. 10, a control chart approach for cash management
19	Section 19.5, stock market indexes
21	Sections 21.7 and 21.8 on mean and variance trade-off analysis and the mean and variance method for capital budgeting decisions; Appendices 2, 3, and 4 of Chap. 21 on the graphical derivation of the capital market line, present value and net present value, and derivation of the standard deviation for NPV

Alternative Three: The objective of the third approach is to show students how calculus can be used in statistical analysis. To achieve this goal, the instructor can try to cover all optional sections and as many of the technical footnotes and appendixes as possible. To do this, of course, the instructor may have to skip many application examples, such as the finance applications discussed in Alternative Two.

Supplementary Materials

Study Guide, by Ahyee Lee, Monmouth College, and Ronald L. Moy, St. John's University. This fine workbook encourages learning by doing. Each chapter begins with a section describing the basic import of each chapter in intuitive terms. Then, the student goes on to a formal review of the chapter and several worked-out problems that show in detail how the solution is obtained. A variety of multiple-choice, true-false, and open-ended questions and problems follows, and finally a brief sample test is given. All answers are included at the end of each chapter.

M1N1TAB Manual, by John C. Lee, University of Illinois. This manual, keyed to the text chapter by chapter, is designed to help students use M1NITAB throughout the course. Each chapter includes a variety of specific applications and ends with both a statistical summary and a summary of MINITAB commands.

Data Sets. A wide variety of macroeconomic, financial, and accounting data is available on computer disks to facilitate student practice. A complete listing of these data sets is given at the end of this book. The disks themselves are free of charge.

Instructor's Guide. The three main parts of the Instructor's Guide are the Overview and Objectives by Cheng F. Lee; the complete Solutions to the text problems by Ahyee Lee and Ronald L. Moy; and the Test Bank, with more than 1,000 multiple-choice and true-false problems, by Alice C. Lee, University of Pennsylvania. Most instructors will find the Instructor's Guide indispensable.

Computerized Testing Program. With the Test Bank on disk for either IBM or Macintosh computers, instructors can select, rearrange, edit, or add problems as they wish.

New Jersey, USA Cheng-Few Lee

Acknowledgments

I am very grateful to my colleagues across the country who have contributed to the development of this book. In particular, I would like to thank Kent Becker, Temple University, and Edward L. Bubnys, Suffolk University, who not only reviewed parts of the manuscript but also class-tested several chapters; John Burr, Mobil Oil Company; H. H. Liao, my research assistant at Rutgers; D. Y. Huang and C. C. Young, both of National Taiwan University; and Kimberly Catucci, my editorial assistant.

I am also indebted to many other people who reviewed all or part of the manuscript:

Richard T. Baillie
Michigan State University

Abdul Basti
Northern Illinois University

Philip Bobko
Rutgers University

Warren Boe
University of Iowa

Y. C. Chang
University of Notre Dame

Shaw K. Chen
The University of Rhode Island

Whewon Cho
Tennessee Technological University

Daniel S. Christiansen
Portland State University

James S. Ford
University of Southern California

Supriya Lahiri
University of Lowell

Leonard Lardaro
The University of Rhode Island

Ahyee Lee
Monmouth College

Keh Shin Lii
University of California

Pi-Erh Lin
Florida State University

Chao-nan Liu
Trenton State College

Tom Mathew
The Troy State University in Montgomery

Richard McGowan
Boston College

Ronald L. Moy
St. John's University

(continued)

(continued)

Mel H. Friedman Kean College	Hassan Pourbabaee University of Central Oklahoma
R. A. Holmes Simon Fraser University	Jean D. Powers The Ohio State University
James Freeland Horrell University of Oklahoma	William E. Stein Texas A&M University
Der Ann Hsu University of Wisconsin-Milwaukee	William Wei Temple University
Dongcheol Kim Rutgers University	Jeffrey M. Wooldridge Massachusetts Institute of Technology
Bharat Kolluri University of Hartford	Gili Yen National Central University, Taiwan

Not least, I would like to thank and salute my family—my wife, Schwinne, for her good humor and patience, and my two children, John and Alice, whose contributions are described elsewhere in this preface.

Cheng F. Lee

Brief Contents

Contents

Part II Probability and Important Distributions

Part I
Introduction and Descriptive Statistics

Part I of this book describes how statistical data can be effectively presented. The effective presentation of data is often very important, whether the presentation itself is a final goal or is to be used as background for further analysis and inference.

Chapter 1 discusses the role of statistics and introduces the basic concepts of descriptive, inferential, deductive, and inductive statistics. Chapter 2 covers data collection and the presentation of data in tables and/or graphs (charts). Chapter 3 discusses how data sets can be organized in a frequency distribution. Finally, in Chapter 4, important statistical measures of various data characteristics are developed and presented; these measures are then used in statistical analyses.

The examples in Part I deal with macroeconomic data, financial ratios, and the rates of return on shares of stock. Other, related topics are also discussed.

Chapter 1
Introduction

Chapter Outline

Key Terms

Statistics	Hypothesis testing
Data	Deduction
Descriptive statistics	Deductive statistical analysis
Inferential statistics	Induction
Population	Inductive statistical analysis
Sample	

1.1 The Role of Statistics in Business and Economics

Statistics is a body of knowledge that is useful for collecting, organizing, presenting, analyzing, and interpreting *data* (collections of any number of related observations) and numerical facts. Applied statistical analysis helps business managers and economic planners formulate management policy and make business decisions more effectively. And statistics is an important tool for students of business and economics. Indeed, business and economic statistics has become one of the most important courses in business education, because a background in applied statistics is a key ingredient in understanding accounting, economics, finance, marketing, production, organizational behavior, and other business courses.

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_1,

We may not realize it, but we deal with and interpret statistics every day. For example, the Dow Jones Industrial Average (DJIA) is the best-known and most widely watched indicator of the direction in which stock market values are heading. When people say, "The market was up 12 points today," they are probably referring to the DJIA. This single statistic summarizes stock prices of 30 large companies. Rather than listing the prices at which all of the approximately 2,000 stocks traded on the New York Stock Exchange are currently selling, analysts and reporters often cite this one number as a measure of overall market performance.

Let's take another example. Before elections, the media sometimes present surveys of voter preference in which a sample of voters instead of the whole population of voters is asked about candidate preferences. The media usually give the results of the poll and then state the possible margin of error. A margin of error of 3 % means that the actual extent of a candidate's popular support may differ from the poll results by as much as 3 % points in either direction ("plus or minus"). Anyone who conducts a survey must understand statistics in order to make such decisions as how many people to contact, how to word the survey, and how to calculate the potential margin of error.

In business and industry, managers frequently use statistics to help them make better decisions. A shoe manufacturer, for instance, needs to produce a forecast of future sales in order to decide whether to expand production. Sales forecasts provide statistical guidance in most business decision making.

On a broader scale, the government publishes a variety of data on the health of the economy. Some of the most popular measures are the gross national product (GNP), the index of leading economic indicators, the unemployment rate, the money supply, and the consumer price index (CPI). All these measures are statistics that are used to summarize the general state of the economy. And, of course, business, government, and academic economists use statistical methods to try to *predict* these macroeconomic and other variables.

The following additional examples are presented to show that the use of statistics is widespread not only in business and economics but in everyday life as well.

Example 1.1 TV Show Ratings. Television executives and advertisers use the ratings provided by A. C. Nielsen to determine which television shows are the most popular. The Nielsen organization regularly surveys a sample of television viewers in the United States about their viewing habits. Their responses are then used to draw conclusions about the viewing habits of the entire US population.

Example 1.2 ABC-GPA. In order to assign letter grades at the end of the semester, a teacher may calculate each student's grade point average to determine how well that student has performed in the class. In so doing, the teacher is calculating the mean or average of a series of grade points. The teacher might also want to know how widely dispersed the scores are across students in that class. In Chap. 4, we will discuss measures that describe the dispersion, or spread, of a group of data.

Example 1.3 One, Two, Three, "Fore!". To improve their golf scores, golfers often compute the average distance they can hit a ball with each golf club. These golfers

then use the mean of a series of measurements to select the best club and thus fine-tune their game.

Example 1.4 Health Benefits of Oat Bran. To determine the health benefits of eating oat bran, a doctor who has access to a large database to which many physicians have contributed compares the average cholesterol level of people who eat oat bran with that of similar people who don't eat oat bran. The doctor is using statistics to evaluate the health benefits of different diets.

Example 1.5 Fertilizer Choice and Plant Growth Rate. Refusing to accept on blind faith the advertising claims of either supplier, a farmer compares the average growth of plants fed with fertilizer A with the average growth of plants fed with fertilizer B to determine which fertilizer is more effective.

1.2 Descriptive Versus Inferential Statistics

Having gotten a feel for the use of statistics by looking at several illustrations, we can now refine our definition of the term. *Statistics* is the collection, presentation, and summary of numerical information in such a way that the data can be easily interpreted.

There are two basic types of statistics: descriptive and inferential. *Descriptive statistics* deals with the presentation and organization of data. Measures of central tendency, such as the mean and median, and measures of dispersion, such as the standard deviation and range, are descriptive statistics. These types of statistics summarize numerical information. For example, a teacher who calculates the mean, median, range, and standard deviation of a set of exam scores is using descriptive statistics. Descriptive statistics is the subject of the first part of this book.

The following are examples of the use (or misuse) of descriptive statistics.

Example 1.6 Baseball Players' Batting Averages. Descriptive statistics can be used to provide a point of reference. The batting averages of baseball players are commonly reported in the newspapers, but to people unfamiliar with baseball, these numbers may be misleading. For example, Wade Boggs of the Boston Red Sox hit .366 in 1988; that is, he got a hit in almost 37 % of his official at bats. Because he was unsuccessful over 63 % of the time, however, a person with little knowledge of baseball might conclude that Boggs is an inferior hitter. Comparing Boggs's average to the mean batting average of all players in the same year, which was .285, reveals that Boggs is among the best hitters.

Example 1. 7 Monthly Unemployment Rates. Graphical statistical analysis can be used to summarize small amounts of information. Figure 1.1 displays the US unemployment rates for each month from January 2001 to July 2011. It shows, for instance, that the unemployment rates for December 2001, December 2005, and December 2010 were 5.7 %, 4.9 %, and 9.4 %, respectively.

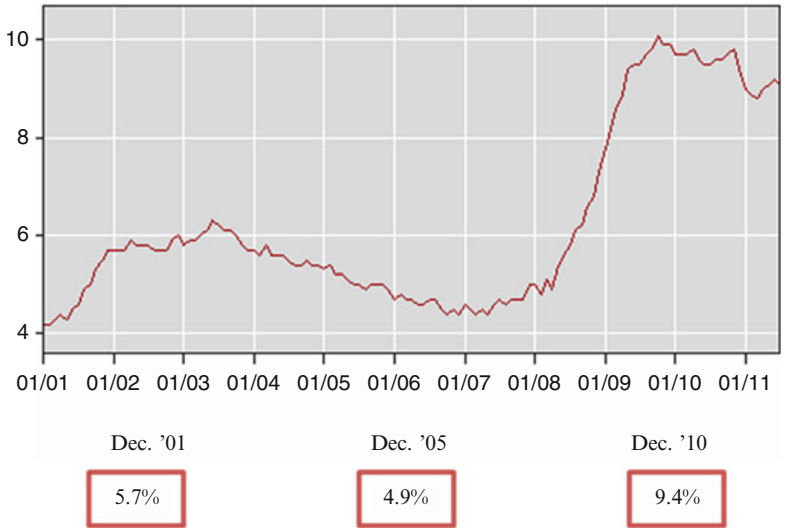


Fig. 1.1 Monthly unemployment rate for the United States (January 2001–July 2011) (*Source*: Bureau of Labor Statistics)

Example 1.8 Comparison of Male and Female Earnings. Descriptive statistics can also be used to compare different groups of data. For example, the mean earnings of full-time working men of different age groups who have had a 4-year college education can be compared with the mean earnings of full-time working women of the same age groups with the same educational background to see whether any differences exist between their earnings. Drawing on 1990 Bureau of the Census data, the *Home News* of central New Jersey used the graph reproduced in Fig. 1.2 to show that mean earnings for full-time working men are higher than those for full-time working women. This figure also shows that the pay gap between full-time working men and women is wider in older age groups. A college-educated woman between the ages of 18 and 24 earns an average of 92 cents for every dollar earned by a man of the same age and educational background. The gap widens steadily as we look at older age groups. Between ages 55 and 64, the average female worker in 1991 was making only 54 cents for every dollar earned by a man of like age and education.[1]

Example 1.9 Returns on Stocks and Bonds. A financial analyst computes financial returns on stocks, corporate bonds, and government bonds to compare their performance during, say, the past 20 years. Because the analyst is collecting and summarizing data, we say that he or she is using descriptive statistics.

Example 1.10 Pitfalls of Comparing the Earnings of Males and Females. We must always be careful when interpreting descriptive statistics. For example, it is sometimes noted that, on average, women earn 70 cents for each dollar that men earn.

[1] This graph is called a bar chart. Bar charts will be discussed in detail in Chap. 2.

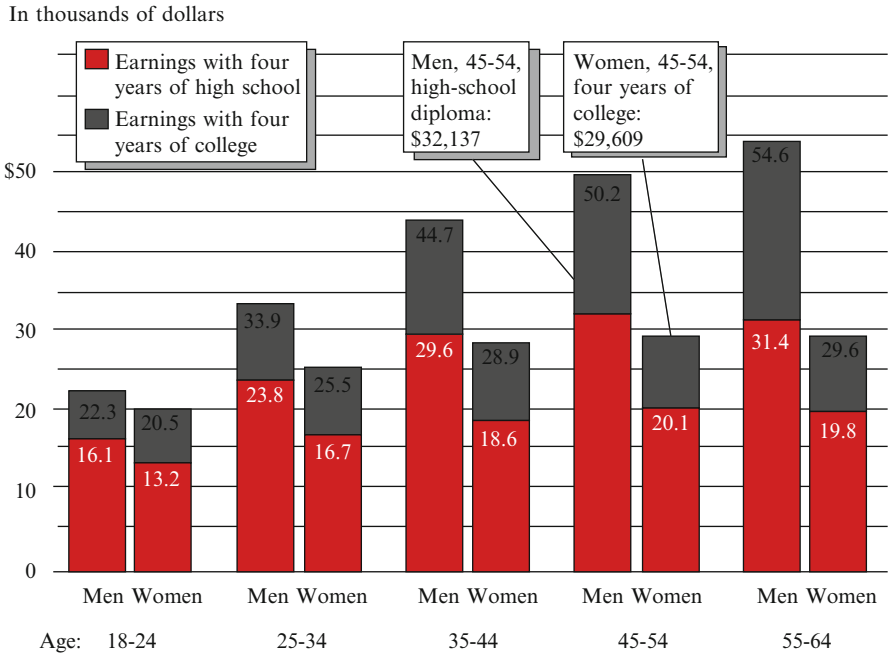


Fig. 1.2 Mean earnings for full-time working men and women, by age and education (*Source*: *Home News*, November 14, 1991. Reprinted by permission of the Associated Press)

That is, the mean earnings for women are compared to the mean earnings for men to suggest that women experience wage discrimination. These descriptive statistics, however, may not tell the whole story. Differences in earnings may result from different occupational choices (which are perhaps influenced by social perceptions of the role women play), from different educational levels and choices (such as which subject to major in), or from career interruptions (which women may experience when choosing to leave their jobs to raise a family). Attributing wage differences entirely to discrimination, then, is an example of the misuse of statistics.[2]

Inferential statistics deals with the use of sample data to infer, or reach, general conclusions about a much larger population. In statistics, we define a *population* as the entire group of individuals we are interested in studying. A *sample* is any subset of such a population. In the election example presented earlier, the pollsters took a sample because it would have been too expensive and time-consuming to contact every voter. This is an example of the use of inferential statistics, because conclusions about a population were made on the basis of sample information. There might be differences between the characteristics of the actual population and the information gained from a sample, so errors can result. Inferential statistics—and in particular

[2] Whether or not career interruptions should be a factor is a matter of debate among behavioral scientists.

hypothesis testing, in which sample information is used to test a hypothesis about a population—is the subject of another major part of this book. Here we will merely look at several examples of inferential statistics.

Example 1.11 Sampling Survey of Residents' Voting Decision. To obtain information on how residents will vote, the *Jericho Clarion* takes a sample and asks the people selected as part of the sample for whom they will vote. This newspaper is using inferential statistics because it is inferring, from a sample, information about a larger population. Again, the newspaper samples the population rather than contacting all its members, because taking a sample is a lot cheaper and less time-consuming.

Example 1.12 Unemployment Rate. The federal government releases information on the unemployment rate every month, which has been discussed in Example 1.7. To arrive at this figure, it samples households across the United States to determine the employment status of the members of those households. Extrapolating from the sample results to the general population is an example of applying inferential statistics.

Example 1.13 Quality Control via Sampling Data. Suppose a production manager of Ford Motor Company compares two samples of a piston produced by different methods to find out whether the two methods result in different fractions of defective units. This production manager takes a sample of 100 pistons produced by one method and checks to see how many are defective and then compares this number to the number of defectives generated by the second production method. One hypothesis is that the number of defectives from the two methods is equal; an alternative hypothesis is that they are not. Inferential statistics can be used here to determine whether the proportions are sufficiently different for the first of these hypotheses to be rejected. This cost-conscious manager has taken a sample to gain information on a much larger quantity.

Example 1.14 A Record Drop in Stock Prices. Statistics is used to summarize the performance of the stock market on a given day. The Dow Jones Industrial Average, an average of the stocks of 30 major firms traded on the New York Stock Exchange, is used as a barometer of the performance of the overall stock market. Other indexes, such as Standard & Poor's 500 Composite Index (S&P 500), the Value Line Index, and the American Stock Exchange Index, are also calculated to generate summary measures of stock market performance. Each of these measures is derived through inferential statistics, because a sample is used to provide representative—though incomplete—information about the stock market at large.[3] For example, the Dow Jones Industrial Average dropped 519.83 points on August 10, 2011. It was the ninth largest point drop in history, as indicated in Table 1.1.

[3] Such is not the case, however, with the NYSE Composite Index. This index is a weighted average (mean) of *all* the firms on the NYSE and is thus the value of a *population* characteristic of a population.

Table 1.1 Ten largest Dow Jones drops (10/27/97–08/10/11)

777.68 pts. to 10,365.45 (7.5 %)	Sep. 29, 2008
733.08 pts. to 8,577.91 (8.5 %)	Oct. 15, 2008
684.81 pts. to 8,920.70 (7.7 %)	Sep. 17, 2001
679.95 pts. to 8,149.09 (8.3 %)	Dec. 1, 2008
678.91 pts. to 8,579.19 (7.9 %)	Oct. 9, 2008
634.76 pts. to 10,809.85 (5.9 %)	Aug. 8, 2011
617.77 pts. to 10,305.78 (6.0 %)	April 14, 2000
554.26 pts. to 7,616.14 (7.7 %)	Oct. 27, 1997
519.83 pts. to 10,719.94 (4.8 %)	Aug. 10, 2011
514.45 pts. to 8,519.21 (6.0 %)	Oct. 22, 2008

Source: Wikipedia

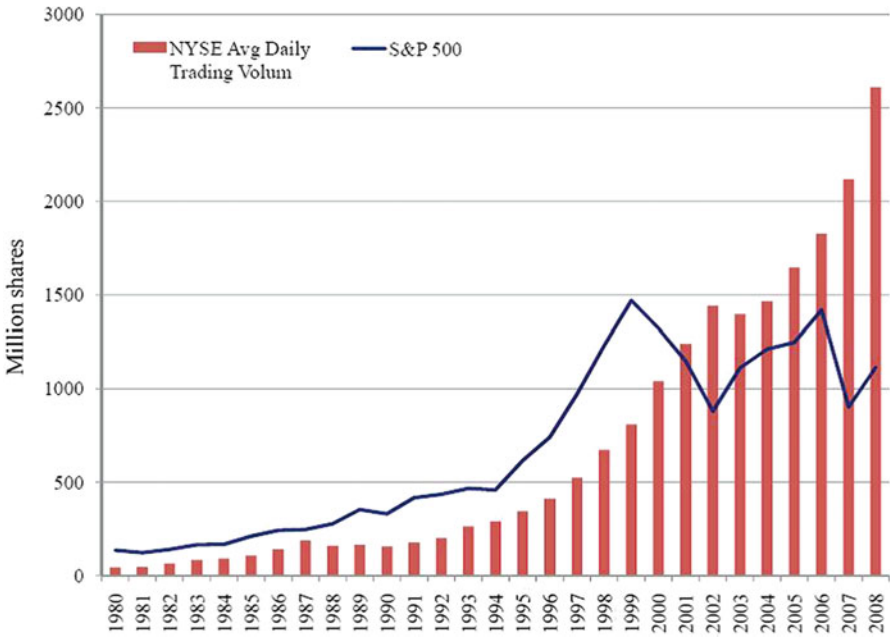


Fig. 1.3 NYSE average daily trading volumes and Standard & Poor's 500 Composite Index (S&P500) (*Source*: NYSE Statistics Archive in NYSE Euronext website)

Example 1.15 Average Daily Trading Volumes. The growth in the number of financial instruments, and in the volume of trade in these instruments, today offers both a timely, crucial, and apropos challenge to current market activities. For example, the amount of trading in New York Stock Exchange has experienced growth, as shown in Fig. 1.3. Secondary markets such as New York Stock Exchange involve already existing issues that are traded among investors. In this case, the instruments are traded between the current investors and the potential investors in a corporation. The proceeds of the sale do not go to the firm but to the current owners of the security.

1.3 Deductive Versus Inductive Analysis in Statistics

We also encounter another dichotomy in statistical analysis. *Deduction* is the use of general information to draw conclusions about specific cases. For example, probability tells us that if a student is chosen by lottery from a calculus class composed of 60 mathematics majors and 40 business administration majors, then the odds against picking a mathematics majors are 4–6. Thus we can deduce that about 40 % of such single-member samples of the students in this calculus class will be business administration majors. As another example of deduction, consider a firm that learns that 1 % of its auto parts are defective and concludes that in any random sample, 1 % of its parts are therefore going to be defective. The use of probability to determine the chance of obtaining a particular kind of sample result is known as *deductive statistical analysis*.

In Chaps. 5, 6, and 7, we will learn how to apply deductive techniques when we know everything about the population in advance and are concerned with studying the characteristics of the possible samples that may arise from that known population.

Induction involves drawing general conclusions from specific information. In statistics, this means that on the strength of a specific sample, we infer something about a general population. The sample is all that is known; we must determine the uncertain characteristics of the population from the incomplete information available. This kind of statistical analysis is called *inductive statistical analysis*. For example, if 56 % of a sample prefers a particular candidate for a political office, then we can estimate that 56 % of the population prefers this candidate. Of course, our estimate is subject to error, and statistics enables us to calculate the possible error of an estimate. In this example, if the error is 3 % points, it can be inferred that the actual percentage of voters preferring the candidate is 56 % plus or minus 3 %; that is, it is between 53 % and 59 %.

Deductive statistical analysis shows how samples are generated from a population, and inductive statistical analysis shows how samples can be used to infer the characteristics of a population. Inductive and deductive statistical analyses are fully complementary. We must study how samples are generated before we can learn to generalize from a sample.

1.4 Summary

This chapter introduced the concept of statistics by presenting examples from everyday life. We saw that statistics can serve as a fundamental tool for decision making not only in business and economics but also in teaching, sports, medicine, quality control, and politics. Finally, we noted that statistics can be classified as either descriptive or inferential, and we drew the distinction between deductive and inductive analysis in statistics.

In the next chapter, we explain the process of collecting data and discuss how to present these data so that they can be interpreted easily and effectively.

Questions and Problems

1. Briefly explain why learning statistical inference is important for students of business and economics. Give two examples.
2. Define the term *statistics*. What are the two basic types of statistics? Describe them and give examples.
3. Explain the difference between deductive and inductive statistics. Illustrate your answer with examples taken from everyday life.
4. You are assigned by your general manager to examine each of last month's sales transactions. Find their average, find the difference between the highest and lowest sates figures, and construct a chart showing the differences between charge account and cash customers. Is this a problem in descriptive or inferential statistics?
5. Suppose you are dealing with problems in probability and statistical inference. Which is usually the larger value in the problem, the population size or the sample size? Why?
6. State which type of statistical problem (deductive or inductive) makes each of the following assumptions.

   (a) You know what the population characteristics are.
   (b) You know what the sample characteristics are.

7. When a cosmetic manufacturer tests the market to determine how many women will buy eyeliner that has been tested for safety without subjecting animals to injury, is it involved in a descriptive statistics problem or an inferential statistics problem? Explain your answer.
8. As controller of the Hamby Corporation, you are directed by the chairman of the board to investigate the problem of overspending by employees who have expense accounts. You ask the accounting department to provide you with records of the number of dollars spent by each of 25 top employees during the past month. The following record is provided:

$292	$494	$600	$807	$535
435	870	725	299	602
322	397	390	420	469
712	520	575	670	723
560	298	472	905	305

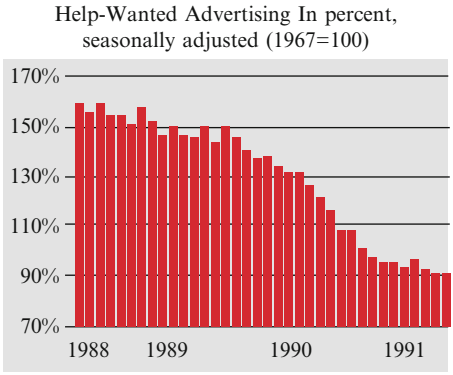
   The question the board of directors wanted to be answered is "How many of our 25 top executives spent more than $600 last month?" What will be your answer? Are you dealing with descriptive statistics or inferential statistics?
9. A teacher has just given an algebra exam. What are some of the statistics she could compute?

10. Suppose the teacher in question 9 is teaching four algebra classes. She would like to predict the average course grade of all her students from only the midterm scores. Should she use inferential or descriptive statistics?
11. Suppose the teacher in question 10 would like to predict the average grade of all her students by using only the midterm scores from one class. Should she use inferential or descriptive statistics?
12. A bullet manufacturer would like to keep the number of duds (bullets that won't fire) to a maximum of 2 per box of 100. Should inferential or descriptive statistics be used to decide this issue? Why?
13. Explain why election pollsters use inferential statistics rather than descriptive statistics to predict the outcome of an election.
14. Explain whether each of the following was arrived at via descriptive or inferential statistics.

    (a) Ted Williams' lifetime batting average
    (b) The number of people watching the Super Bowl, based on A. C. Nielsen's ratings
    (c) The number of people who favor teacher-led prayer in school, based on a survey of churchgoers
    (d) The average rate of return for IBM stock over the last 10 years

15. Suppose you are interested in purchasing AT&T stock. You know that AT&T stock has had an average rate of return of 8 % over the last 5 years. Explain how you could use descriptive statistics to help you decide whether to purchase AT&T.
16. A popular commercial claims that "Four out of five dentists prefer sugarless gum." Is this conclusion drawn from a sample or a population?
17. The most commonly reported indicator of stock market performance is the Dow Jones Industrial Average. Explain whether the firms whose share price are included in the DJIA represent a sample or a population.
18. Use the information given in the accompanying figure to answer the following questions.

    (a) Which month has the greatest amount of help-wanted advertising?
    (b) How did help-wanted advertising fluctuate during the period of October 1988 through September 1991?



Source: Wall Street Journal, November 7, 1991, p. A1

19. Suppose a Gallup poll is to be conducted to predict the outcome of the 1992 presidential election. Should the pollsters survey a sample or a population?
20. If managers at Weight Watchers are interested in the average number of pounds that people on Weight Watchers diets lose, should they use a sample or the population to find out?
21. Suppose that in question 20, Weight Watchers uses a sample. Does this represent the use of inferential or descriptive statistics? How would your answer change if Weight Watchers used the population?
22. The following list gives the seven highest-paid baseball players for 1992 and their salaries.

Dwight Gooden	$5,166,666
George Brett	4,700,000
Roger Clemens	4,300,000
Will Clark	4,250,000
Andy Van Slyke	4,250,000
Darryl Strawberry	4,050,000
Fred McGriff	4,000,000

(a) Does this list represent a sample or population of baseball players' salaries?
(b) If you were the agent for a top baseball star, how could you use the foregoing information?

23. Suppose Greg Norman has the following scores in his last eight rounds of golf: 71, 68, 64, 73, 69, 62, 75, 69.

(a) If Greg computed his average score over these eight rounds, would he be computing a descriptive or an inferential statistic?
(b) If Greg used these scores to predict his overall scoring average for the 1993 season, would he be using descriptive or inferential statistics?

24. Suppose a real estate broker in Albany, New York, is interested in the average price of a home in a development comprising 100 homes.

    (a) If she uses 12 homes to predict the average price of all 100 homes, is she using inferential or descriptive statistics?
    (b) If she uses all 100 homes, is she using inferential or descriptive statistics?

25. J. D. Power is a consulting firm that assesses consumer satisfaction for the auto industry. Do you think this company uses a sample or the population to conduct its survey?
26. Using any newspaper of your choice, find examples of statistics from the following sections:

    (a) Sports section
    (b) Business section
    (c) Entertainment section

27. Are the statistics you found in answering question 26 inferential or descriptive statistics?
28. The owner of a factory regularly requests a graphical summary of all employees' salaries. The graphical summary of salaries is an example of descriptive statistics inferential statistics?
29. A manager asked 50 employees in a company of their ages. On the basis of this information, the manager states that the average age of all the employees in the company is 39 years. The statement of the manager is an example of descriptive statistics inferential statistics.
30. Refer to Table 2.2, in which annual macroeconomic data including GDP, CPI, 3-month T-bill rate, prime rate, private consumption, private investment, net exports, and government expenditures from 1960 to 2009 are given. Answer the following questions.

    (a) How many observations are in the data set?
    (b) How many variables are in the data set?

31. Refer to Table 2.2, which of the variables are qualitative and which are quantitative variables?

Chapter 2
Data Collection and Presentation

Chapter Outline

Key Terms

Primary data	Pie charts
Secondary data	Balance sheet
Census	Income statement
Sample	Assets
Random error	Liabilities
Systematic error	Net worth
Time-series graph	Liquidity ratios
Line chart	Leverage ratios
Component-parts line chart	Activity ratios
Component-parts line graph	Profitability ratios
Bar charts	Market value ratios

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_2, © Springer Science+Business Media New York 2013

2.1 Introduction

The collection, organization, and presentation of data are basic background material for learning descriptive and inferential statistics and their applications. In this chapter, we first discuss sources of data and methods of collecting them. Then we explore in detail the presentation of data in tables and graphs. Finally, we use both accounting and financial data to show how the statistical techniques discussed in this chapter can be used to analyze the financial condition of a firm and to analyze the recent deterioration of the financial health of the US banking industry. In addition, we use a pie chart to examine how Congress voted on the Gulf Resolution in 1991.

2.2 Data Collection

After identifying a research problem and selecting the appropriate statistical methodology, researchers must collect the data that they will then go on to analyze. There are two sources of data: primary and secondary sources. *Primary data* are data collected specifically for the study in question. Primary data may be collected by methods such as personal investigation or mail questionnaires. In contrast, *secondary data* were not originally collected for the specific purpose of the study at hand but rather for some other purpose. Examples of secondary sources used in finance and accounting include the *Wall Street Journal, Barron's, Value Line Investment Survey, Financial Times*, and company annual reports. Secondary sources used in marketing include sales reports and other publications. Although the data provided in these publications can be used in statistical analysis, they were not specifically collected for that use in any particular study.

Example 2.1 Primary and Secondary Sources of Data. Let us consider the following cases and then characterize each data source as primary or secondary:

1. (Finance) To determine whether airline deregulation has increased the return and risk of stocks issued by firms in the industry, a researcher collects stock data from the *Wall Street Journal* and the Compustat database. (The Compustat database contains accounting and financial information for many firms.)
2. (Production) To determine whether ball bearings meet measurement specifications, a production engineer examines a sample of 100 bearings.
3. (Marketing) Before introducing a hamburger made with a new recipe, a firm gives 25 customers the new hamburger and asks them on a questionnaire to rate the hamburger in various categories.
4. (Political science) A candidate for political office has staff members call 1,000 voters to determine what candidate they prefer in an upcoming election.
5. (Marketing) A marketing firm looks up, in *Consumer Reports,* the demand for different types of cars in the United States.

6. (Economics) An economist collects data on unemployment from a Department of Labor report.
7. (Accounting) An accountant uses sampling techniques to audit a firm's accounts receivable or its inventory account.
8. (Economics) The staff from the Department of Labor uses a survey to estimate the current unemployment rate in the United States.

The cases numbered 1, 5, and 6 illustrate the use of secondary sources; these researchers relied on existing data sets. The remainder involve primary sources because the data involved were generated specifically for that study.

The main advantage of primary data is that the investigator directly controls how the data are collected; therefore, he or she can ensure that the information is relevant to the problem at hand. For example, the investigator can design the questionnaires and surveys to elicit the most relevant information. The disadvantage of this method is that developing appropriate surveys or questionnaires requires considerable time, money, and experience. In addition, mail questionnaires are usually plagued by a low response rate. What response rate is acceptable varies with context and with other factors. A response rate of 50 % is often considered acceptable, but it is rarely achieved with mail questionnaires.

Fortunately, there are many good secondary sources of information in business, economics, and finance. Financial information such as stock prices and accounting data is easy to locate but tedious to organize. As an alternative, databases such as Compustat and CRSP (Center for Research on Securities Prices) tapes can be used. Economic data can be found in many government publications, such as the *Federal Reserve Bulletin,* the *Economic Report of the President*, and the *Statistical Abstract of the United States.* In addition, macroeconomic variables are found in databases such as that of Data Resources. Of course, not all secondary sources are unimpeachable. Possible problems include outdated data, the restrictive definitions used, and unreliability of the source.

A sample or a census may be taken from either primary or secondary data. A *census* contains information on *all* members of the population; a *sample* contains observations from a *subset* of it. A census of primary data results, for example, from the polling of all voters in a city to determine their preference for mayor. If a subset of voters in the city is asked about their preference for mayor, a sample of primary data results. These are both examples of using primary data because the data are collected for purposes of the study that is under way.

If a researcher records the prices of all the securities traded on the New York Stock Exchange for 1 day as they are listed in the *Wall Street Journal*, he or she is taking a census from a secondary source. However, if he or she takes a subset from the population—say, every fifth price—he or she is developing a sample of secondary data. Note that taking stock prices from the newspaper is an example of using secondary data because the data were not collected specifically for the study.

Given that the purpose of taking a sample is to gain information on a population, why do we not take a census every time we need information? The first reason is the high cost of taking a census. It would be extremely expensive for a pollster who

wanted information on the outcome of a presidential election to contact all the registered voters in the country. Of course, the costs of obtaining the names of voters, hiring people to conduct the survey, performing computer analysis, and carrying out research must also be incurred when taking a sample, but because the sample is usually much smaller than the population, these costs are substantially reduced.

For example, to determine Illinois voters' preferences in the 1988 presidential election, the *Chicago Tribune* sampled 766 Illinois residents who said they would vote in the election. Obviously, sampling was cheaper than contacting all Illinois adults. The poll was accurate to within five percentage points, which is an acceptable margin of error. In Chaps. 8 and 20, we will return to the topic of calculating the error in sampling.

The second advantage of sampling is accuracy. Because fewer people are contacted in a sample, the interviewers can allot more time to each respondent. In addition, the need for fewer workers to conduct the study may make it possible to select and train a more highly qualified staff of researchers. This, in turn, may result in a study of higher quality.

Another problem in taking a large census is the time involved. For example, suppose it would take at least 2 months for the *Tribune* to contact all the adults in Illinois. If the election were only 1 month away, the poll would not be of any use. In cases where the population is very large and will take a long time to reach, a sample is the more timely method of obtaining information.

This is not to suggest that a sample is always better than a census. A census is appropriate when the population is fairly small. For example, a census would be feasible if you wanted information on how the members of a small high school class intended to vote for student council president because the cost and time of contacting every member of the class would be relatively low. In contrast, a sample is more cost- and time-effective when the population is a city, state, nation, or other large entity.

There are two types of errors that can arise when we are dealing with primary or secondary data. The first is *random error*, which is the difference between the value derived by taking a random sample and the value that would have been obtained by taking a census. This error arises from the random chance of obtaining the specific units that are included in the sample. Happily, random error can be reduced by increasing the sample size, and it can be reduced to zero by taking a census. Random error can also be estimated. Using statistics, the *Chicago Tribune* was able to determine that this poll was subject to random error of plus or minus 5 %. This issue will be discussed in Part III, on sampling and statistical inference.

Systematic error results when there are problems in measurement. Unlike random error, which can occur only in sampling, systematic error can occur in both samples and census. For example, suppose that a basketball coach measures the heights of his players with an imprecise ruler. The resulting error is "systematic": the ruler distorts all measurements equally. As another example, when a researcher uses an incorrect computer program that calculates an arithmetic mean

by dividing by the number of observations plus 5, a systematic error results because the divisor should have been the number of observations.

Let us use the measurement of basketball players' heights to compare random and systematic errors. Suppose the basketball coach selects a sample of five players, measures their heights with a "good" ruler, and finds (by dividing properly) that the mean of the sample is 6 ft 1 in. If the actual average height of all the players is 6 ft 2 in., the mean random error is −1 in. A random error will result. Now suppose the coach uses a ruler that is 2 in. too short. When measuring all the players'(a census), he comes up with a population mean of 6 ft even. In this case, a systematic error of −2 in. results.

2.3 Data Presentation: Tables

All data tables have four elements: a caption, column labels, row labels, and cells. The caption describes the information that is contained in the table. The column labels identify the information in the columns, such as the gross national product, the inflation rate, or the Dow Jones Industrial Average. Examples of row labels include years, dates, and states. A cell is defined by the intersection of a specific row and a specific column.

Example 2.2 Annual CPI, T-Bill Rate, and Prime Rate. To illustrate, Table 2.1 gives some macroeconomic information from 1950 to 2010. The caption is "CPI, T-bill rate, and prime rate (1950–2010)." The row labels are the years 1950–2010. The column labels are CPI (consumer pace index), 3-month T-bill rate, and prime rate. Changes in the consumer price index, the most commonly used indicator of the economy's price level, are a measure of inflation or deflation. (For a more detailed description of the CPI, see Chap. 19.) The 3-month T-bill interest rate is the interest rate that the USA Treasury pays on 91-day debt instruments, and the prime rate is the interest rate that banks charge on loans to their best customers, usually large firms. This table, then, presents macroeconomic information for any year indicated. For example, the CPI for 2010 was 218.1 and the prime rate in 2008 was 5.09 %. The relationship between the CPI and 3-month T-bill rate will be discussed in Chap. 19.

2.4 Data Presentation: Charts and Graphs

It is sometimes said that a picture is worth a thousand words, and nowhere is this statement more true than in the analysis of data. Tables are usually filled with highly specific data that take time to digest. Graphs and charts, though they are often less detailed than tables, have the advantage of presenting data in a more accessible and

Table 2.1 CPI, T-bill rate, and the prime rate (1950–2010)

Year	CPI[a]	3-Month T-bill rate	Prime rate
50	24.1	1.218	2.07
51	26	1.552	2.56
52	26.5	1.766	3
53	26.7	1.931	3.17
54	26.9	0.953	3.05
55	26.8	1.753	3.16
56	27.2	2.658	3.77
57	28.1	3.267	4.2
58	28.9	1.839	3.83
59	29.1	3.405	4.48
60	29.6	2.928	4.82
61	29.9	2.378	4.5
62	30.2	2.778	4.5
63	30.6	3.157	4.5
64	31	3.549	4.5
65	31.5	3.954	4.54
66	32.4	4.881	5.63
67	33.4	4.321	5.61
68	34.8	5.339	6.3
69	36.7	6.677	7.96
70	38.8	6.458	7.91
71	40.5	4.348	5.72
72	41.8	4.071	5.25
73	44.4	7.041	8.03
74	49.3	7.886	10.81
75	53.8	5.838	7.86
76	56.9	4.989	6.84
77	60.6	5.265	6.83
78	65.2	7.221	9.06
79	72.6	10.041	12.67
80	82.4	11.506	15.27
81	90.9	14.029	18.87
82	96.5	10.686	14.86
83	99.6	8.63	10.79
84	103.9	9.58	12.04
85	107.6	7.48	9.93
86	109.6	5.98	8.33
87	113.6	5.82	8.22
88	118.3	6.69	9.32
89	124	8.12	10.87
90	130.7	7.51	10.01
00	172.2	5.85	9.23
01	177.1	3.44	6.91
02	179.9	1.62	4.67
03	184	1.01	4.12
04	188.9	1.38	4.34

(continued)

Table 2.1 (continued)

Year	CPI[a]	3-Month T-bill rate	Prime rate
05	195.3	3.16	6.19
06	201.6	4.73	7.96
07	207.3	4.41	8.05
08	215.3	1.48	5.09
09	214.5	0.16	3.25
10	218.1	0.14	3.25

Source: Economic Report of the President, January 2010
[a]CPI base: 1982–1984 = 100

memorable form. In most graphs and charts, the independent variable is plotted on the horizontal axis (the *x*-axis) and the dependent variable on the vertical axis (the *y*-axis). Frequently, "time" is plotted along the *x*-axis. Such a graph is known as a *time-series graph* because on it, changes in a dependent variable (such as GDP, inflation rate, or stock prices) can be traced over time.

Line charts are constructed by graphing data points and drawing lines to connect the points. Figure 2.1 shows how the rate of return on the S&P 500 and the 3-month T-bill rate have varied over time.[1] The independent variable is the year (ranging from 1990 to 2010), so this is a time-series graph. The dependent variables are often in percentages.

Figure 2.2 is a graph of the components of the gross domestic product (GDP)—personal consumption, government expenditures, private investment, and net exports—over time. This is also a time-series graph because the independent variable is time. It is a *component-parts line chart*. These series have been "deflated" by expressing dollar amounts in constant 2005 dollars. (Chap. 19 discusses the deflated series in further detail.)

Figure 2.2 is also called a *component-parts line graph* because the four parts of the GDP are graphed. The sum of the four components equals the GDP. Using this type of graph makes it possible to show the sources of increases or declines in the GDP. (The data used to generate Fig. 2.2 are found in Table 2.2.)

Bar charts can be used to summarize small amounts of information. Figure 2.3 shows the average annual returns for Tri-Continental Corporation for investment periods of seven different durations ending on September 30, 1991. This figure shows that Tri-Continental has provided investors double-digit returns during a 50-year period.

It also shows that the investment performance of this company was better than that of the Dow Jones Industrial Average (DJIA) and the S&P 500.[2]

[1] T-bill rate data can be found in Table 2.1; rates of return on the S&P 500 can be found in Table 2.4 in Appendix 2 of this chapter. Most of the figures in this book are drawn with the Microsoft Excel PC program. The procedure for using the Excel program to draw these graphs can be found in Appendix 1 of this chapter.

[2] Both the DJIA and the S&P 500 will be discussed in Chap. 19 of this book.

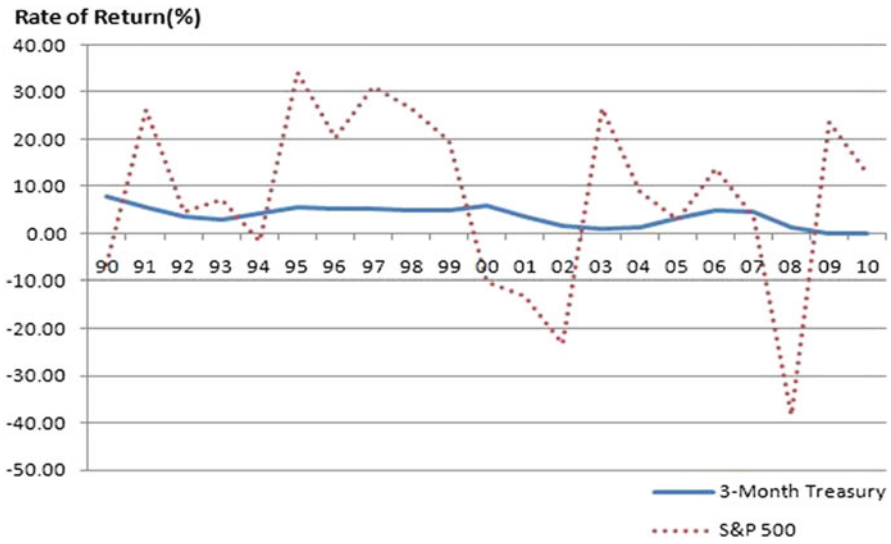


Fig. 2.1 Rates of return on S&P 500 and 3-month T-bills

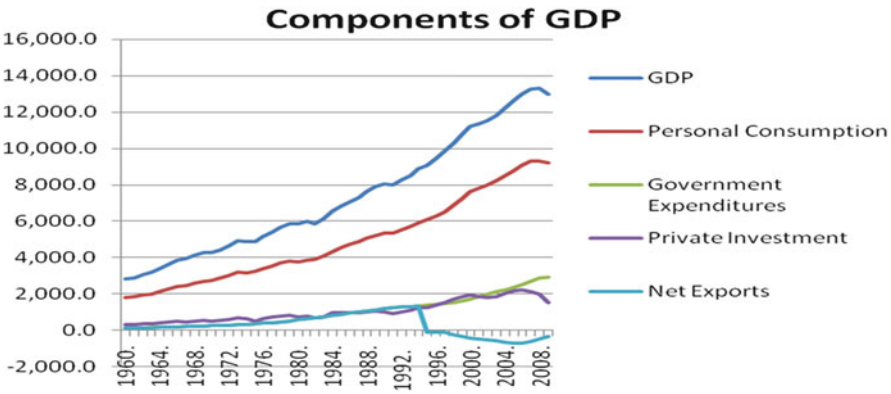


Fig. 2.2 Components of GDP (billions of 2005 dollars)

As this example illustrates, using a bar graph is most appropriate when we are comparing only a few items.

Pie charts are used to show the proportions of component parts that make up a total. Figure 2.4 shows how the US soft drink market was broken down in 1985. The two industry leaders, Coca-Cola and PepsiCo, enjoyed 40 % and 28 % of the market share, respectively. The next four largest firms (Seven-Up, Dr Pepper, Royal

Table 2.2 Annual macroeconomic data 1960–2009 (in 2005 dollars)

Year	GDP[a]	CPI[b]	3-month T-bill rate	Prime rate	Private consumption[a]	Private investment[a]	Net exports[a]	Government expenditures[a]
1960	2,830.9	29.585	2.88	4.82	1,784.4	296.5	111.5	111.5
1961	2,896.9	29.902	2.35	4.50	1,821.2	294.6	119.5	119.5
1962	3,072.4	30.253	2.77	4.50	1,911.2	332.0	130.1	130.1
1963	3,206.7	30.633	3.16	4.50	1,989.9	354.3	136.4	136.4
1964	3,392.3	31.038	3.55	4.50	2,108.4	383.5	143.2	143.2
1965	3,610.1	31.528	3.95	4.54	2,241.8	437.3	151.4	151.4
1966	3,845.3	32.471	4.86	5.63	2,369.0	475.8	171.6	171.6
1967	3,942.5	33.375	4.31	5.63	2,440.0	454.1	192.5	192.5
1968	4,133.4	34.792	5.34	6.31	2,580.7	480.5	209.3	209.3
1969	4,261.8	36.683	6.67	7.95	2,677.4	508.5	221.4	221.4
1970	4,269.9	38.842	6.39	7.91	2,740.2	475.1	233.7	233.7
1971	4,413.3	40.483	4.33	5.72	2,844.6	529.3	246.4	246.4
1972	4,647.7	41.808	4.07	5.25	3,019.5	591.9	263.4	263.4
1973	4,917.0	44.425	7.03	8.02	3,169.1	661.3	281.7	281.7
1974	4,889.9	49.317	7.83	10.80	3,142.8	612.6	317.9	317.9
1975	4,879.5	53.825	5.78	7.86	3,214.1	504.1	357.7	357.7
1976	5,141.3	56.933	4.97	6.84	3,393.1	605.9	383.0	383.0
1977	5,377.7	60.617	5.27	6.82	3,535.9	697.4	414.1	414.1
1978	5,677.6	65.242	7.19	9.06	3,691.8	781.5	453.6	453.6
1979	5,855.0	72.583	10.07	12.67	3,779.5	806.4	500.7	500.7
1980	5,839.0	82.383	11.43	15.27	3,766.2	717.9	566.1	566.1
1981	5,987.2	90.933	14.03	18.87	3,823.3	782.4	627.5	627.5
1982	5,870.9	96.533	10.61	14.86	3,876.7	672.8	680.4	680.4
1983	6,136.2	99.583	8.61	10.79	4,098.3	735.5	733.4	733.4
1984	6,577.1	103.933	9.52	12.04	4,315.6	952.1	796.9	796.9
1985	6,849.3	107.600	7.48	9.93	4,540.4	943.3	878.9	878.9
1986	7,086.5	109.692	5.98	8.33	4,724.5	936.9	949.3	949.3
1987	7,313.3	113.617	5.78	8.20	4,870.3	965.7	999.4	999.4
1988	7,613.9	118.275	6.67	9.32	5,066.6	988.5	1,038.9	1,038.9
1989	7,885.9	123.942	8.11	10.87	5,209.9	1,028.1	1,100.6	1,100.6
1990	8,033.9	130.658	7.49	10.01	5,316.2	993.5	1,181.7	1,181.7
1991	8,015.1	136.167	5.38	8.46	5,324.2	912.7	1,236.1	1,236.1
1992	8,287.1	140.308	3.43	6.25	5,505.7	986.7	1,273.5	1,273.5
1993	8,523.4	144.475	3.00	6.00	5,701.2	1,074.8	1,294.8	1,294.8
1994	8,870.7	148.225	4.25	7.14	5,918.9	1,220.9	1,329.8	1,329.8
1995	9,093.7	152.383	5.49	8.83	6,079.0	1,258.9	−98.8	1,374.0
1996	9,433.9	156.858	5.01	8.27	6,291.2	1,370.3	−110.7	1,421.0
1997	9,854.3	160.525	5.06	8.44	6,523.4	1,540.8	−139.8	1,474.4
1998	10,283.5	163.008	4.78	8.35	6,865.5	1,695.1	−252.6	1,526.1
1999	10,779.8	166.583	4.64	7.99	7,240.9	1,844.3	−356.6	1,631.3
2000	11,226.0	172.192	5.82	9.23	7,608.1	1,970.3	−451.6	1,731.0
2001	11,347.2	177.042	3.39	6.92	7,813.9	1,831.9	−472.1	1,846.4
2002	11,553.0	179.867	1.60	4.68	8,021.9	1,807.0	−548.8	1,983.3
2003	11,840.7	184.000	1.01	4.12	8,247.6	1,871.6	−603.9	2,112.6

(continued)

Table 2.2 (continued)

Year	GDP[a]	CPI[b]	3-month T-bill rate	Prime rate	Private consumption[a]	Private investment[a]	Net exports[a]	Government expenditures[a]
2004	12,263.8	188.908	1.37	4.34	8,532.7	2,058.2	−688.0	2,232.8
2005	12,638.4	195.267	3.15	6.19	8,819.0	2,172.2	−722.7	2,369.9
2006	12,976.2	201.550	4.73	7.96	9,073.5	2,230.4	−729.2	2,518.4
2007	13,254.1	207.335	4.35	8.05	9,313.9	2,146.2	−647.7	2,676.5
2008	13,312.2	215.247	1.37	5.09	9,290.9	1,989.4	−494.3	2,883.2
2009	12,988.7	214.549	0.15	3.25	9,237.3	1,522.8	−353.8	2,933.3

Source: Department of Commerce (Bureau of Economic Analysis), Economic Report of the President, February 2010
[a]Billions of 2005 dollars
[b]CPI base: 1982–1984 = 100

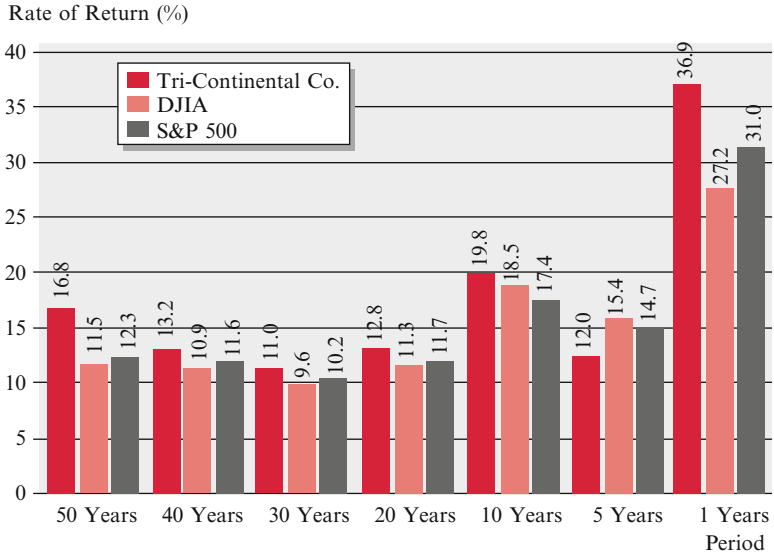


Fig. 2.3 Average annual returns for Tri-Continental Corporation for investment periods of seven different durations ending on September 30, 1991 (*Source: Wall Street Journal*, November 18, 1991. p. C5)

Crown, and Cadbury Schweppes) accounted for 21.8 % of the market, and the remaining 10.2 % of the market was divided among still smaller companies.

2.5 Applications

In the last several sections, we have drawn primarily on macroeconomic data to show how tables and graphs can be used to examine various economic variables. In this section, we will use the same tabular and graphical tools to analyze financial and accounting data that are important in financial analysis and planning. We also will see how Congress voted on the Gulf Resolution.

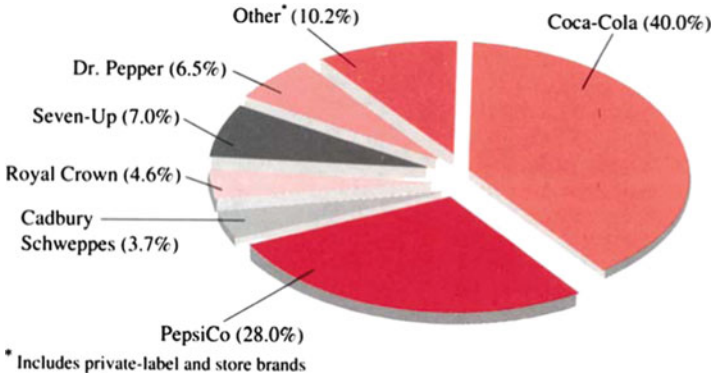


Fig. 2.4 US soft drink market breakdown (1985) (*Data: Beverage Digest*, Montgomery Securities. *Source: Business Week*)

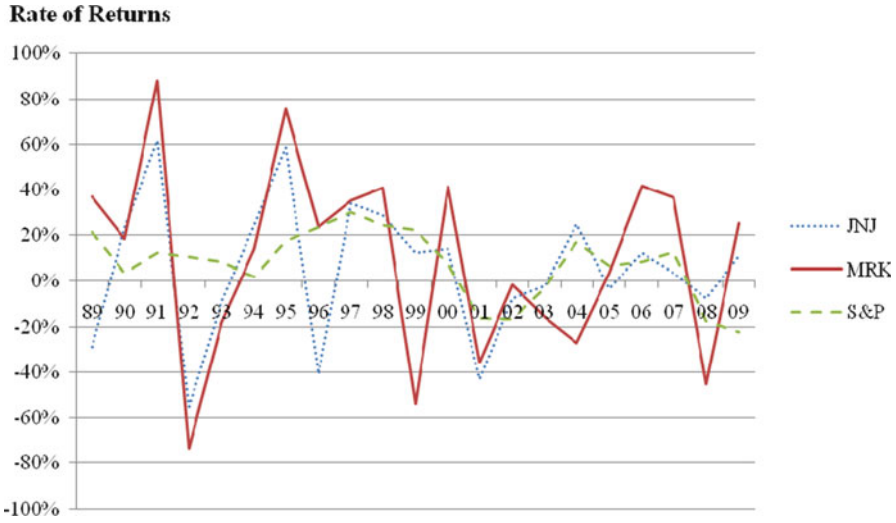


Fig. 2.5 Rates of return for S&P 500, Johnson & Johnson, and Merck

Application 2.1 Analysis of Stock Rates of Return and Market Rates of Return. Stock prices and stock indexes are two familiar measures of stock market performance. In addition to these indicators, percentage rates of return can be calculated to determine how well a particular stock—or the stock market overall—is doing.

Figure 2.5 is a line graph of yearly rates of return for Johnson & Johnson, Merck, and the S&P 500, which, as we have noted, is a market index. The yearly rates of return have been similar for the three. This indicator has fluctuated relatively

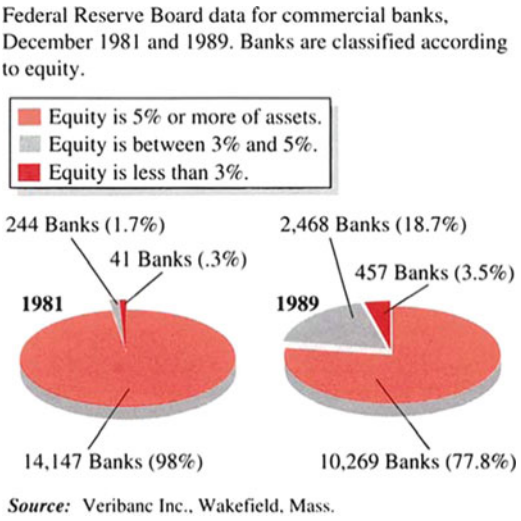


Fig. 2.6 How healthy is your bank? (*Source: Home News*, January 6, 1991. Reprinted by permission of The Associated Press)

similar as well for the Merck stock and Johnson & Johnson stock, while the overall market (as gaged by the S&P 500) has varied least of all.[3]

Application 2.2 Financial Health of the Banking Industry: 1981 Versus 1989. On January 6, 1991, the business section of the *Home News* (a central New Jersey newspaper) printed an Associated Press article that used two pie charts prepared by VERIBANC Inc., a financial rating service. The pie charts, presented in Fig. 2.6, compare the financial condition of the US commercial banks in 1981 to their condition in 1989. These two pie charts show that the percentage of nonproblem banks (those whose equity is 5 % or more of their assets) has fallen from 98 % to 77.8 %, revealing that the probability that depositors are dealing with a problem-plagued bank has increased about 11 times.[4] In view of this deterioration, the article offers the following five tips to anyone shopping for a new financial institution:

1. Determine whether deposits are protected by federal deposit insurance, which covers deposits of up to $100,000.
2. Research any state deposit insurance funds.
3. Investigate the institution's history.
4. Check new reports for the health of specific banks and other industry trends.
5. Ask the bank for its yearly financial statement. Or contact federal bank regulators for the institution's quarterly statement of financial condition and its income statement.

[3] Rates of return for Johnson & Johnson and Merck and market rates of return are analyzed in more detail in Appendix 2.

[4] Eleven times can be calculated as $\frac{1.0-.778}{1.0-.98} = .111$.

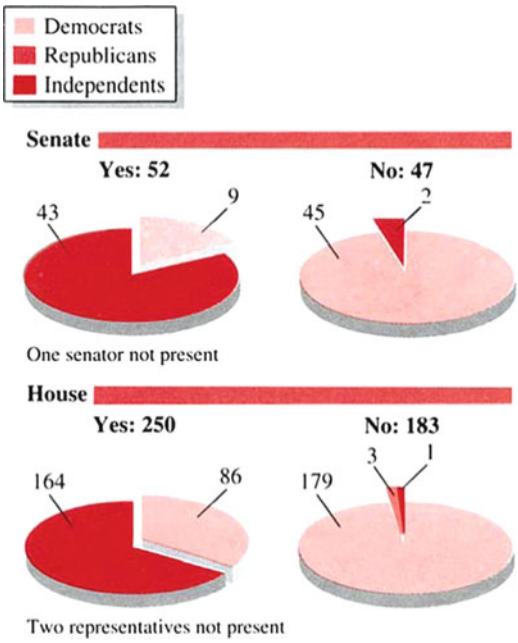


Fig. 2.7 How congress voted on the Gulf Resolution (*Source: Home News*. January 6, 1991. Reprinted by permission of The Associated Press)

Application 2.3 How Congress Voted on the Gulf Resolution. Following a heated debate, Congress voted to grant President Bush the power to go to battle against Iraq if the Iraqis did not withdraw from Kuwait by January 15, 1991.

As indicated in the pie chart in Fig. 2.7, the Senate vote of 52–47 and the House vote of 250–183 authorized President Bush to use military force against Iraq. Among those voting *yes* were 43 Republican senators, 9 Democratic senators, 164 Republicans in the House of Representatives, and 86 Democrats in the House of Representatives; among those voting *no* were 2 Republican senators and 45 Democratic senators. In the House, 3 Republicans, 179 Democrats, and 1 independent voted no. In terms of percentages, about 52.53 % of the Senate and 57.74 % of the House voted to support the Gulf Resolution. One senator and two representatives were not present.

Application 2.4 Bar Charts Reveal How Several Economic Indicators Are Related. In *Time* magazine, November 1991, eight bar charts showed how economic conditions fluctuated during 1989–1991.

To stimulate the economy, policy decision makers at the Fed lowered interest rates. Mortgage rates dropped from 10.32 % in 1989 to 8.76 % in 1991, while auto loan rates dropped from 12.27 % to 11.78 %. Despite these lower interest rates, both housing starts and auto sales experienced surprising declines. Figure 2.8a–h gives a clear picture of these relationships.

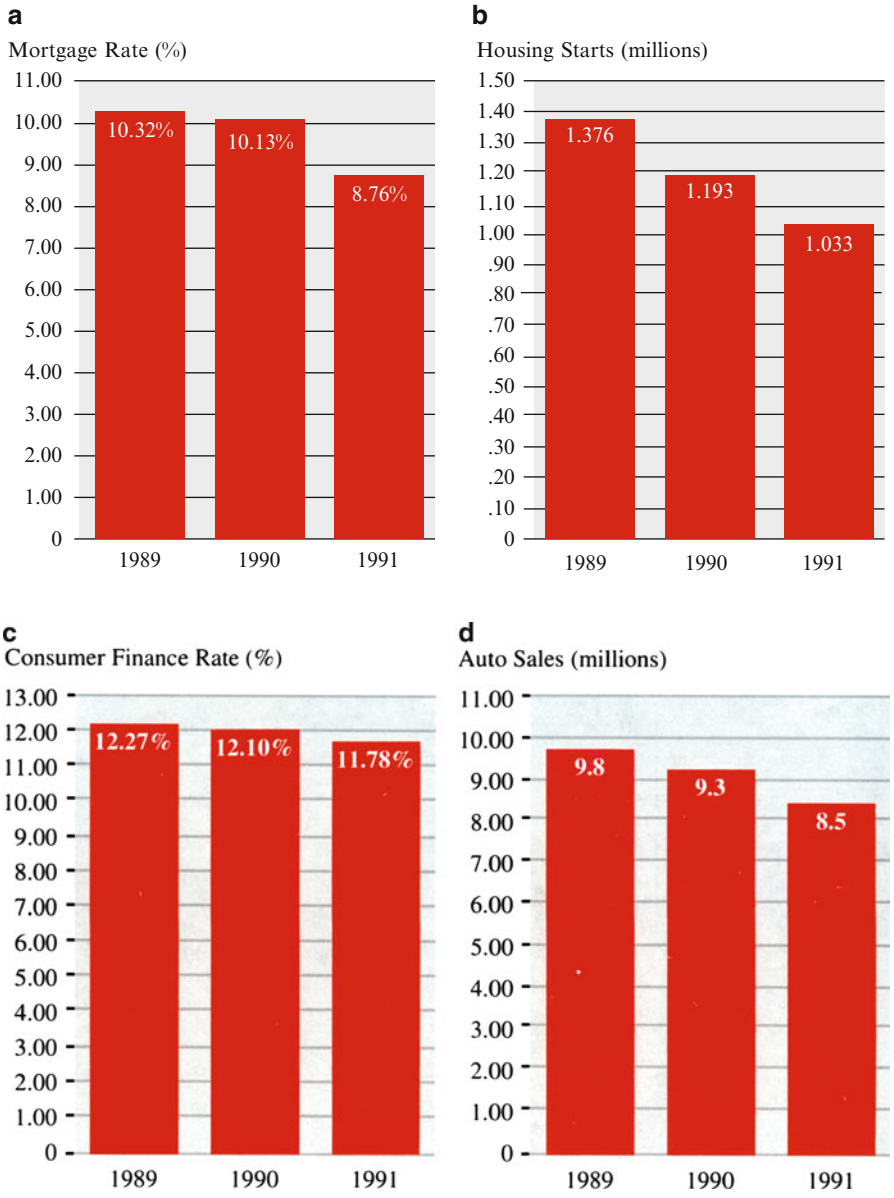


Fig. 2.8 (continued)

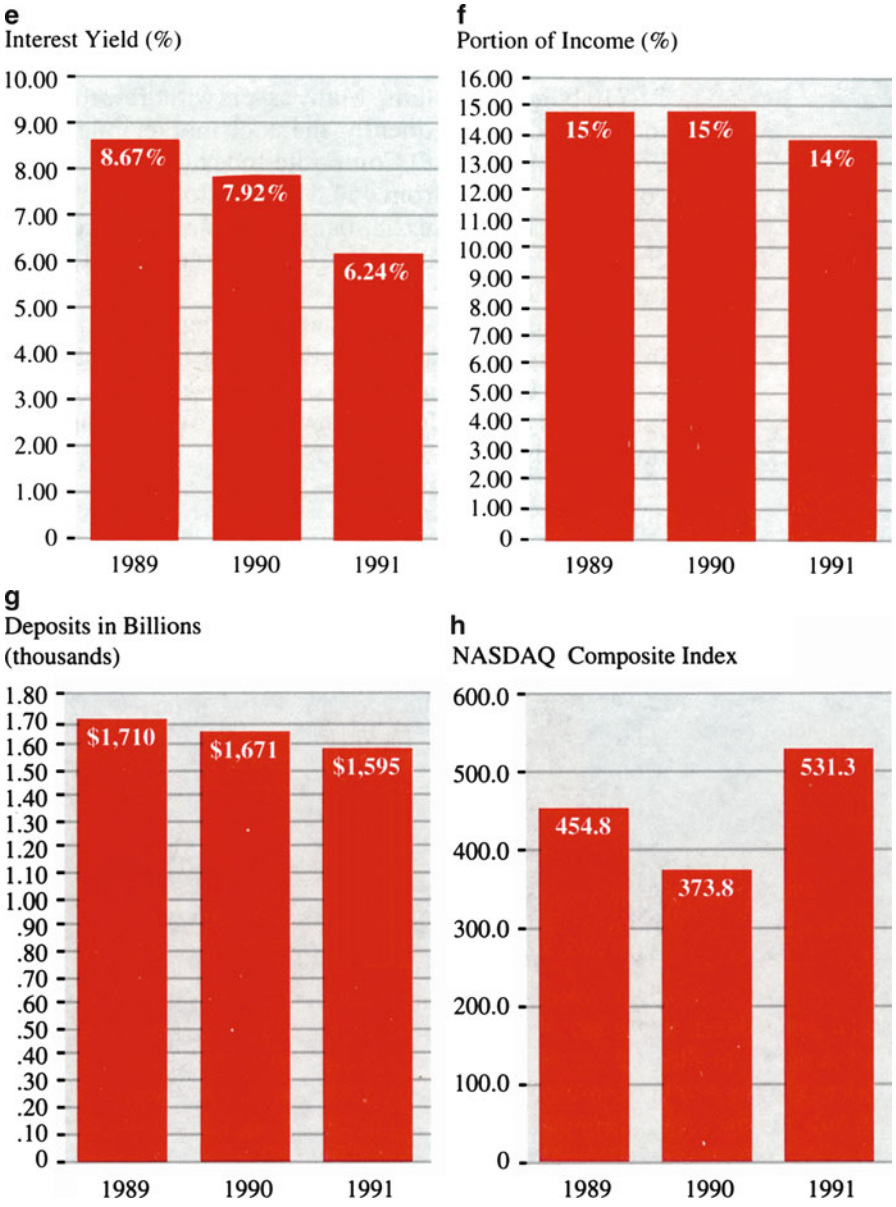


Fig. 2.8 Eight macroeconomic indicators: (**a**) mortgage rates, (**b**) housing starts, (**c**) consumer finance rate for auto loans, (**d**) domestic auto sales, (**e**) 1-year CD interest yield, (**f**) portion of income derived from interest, (**g**) time and savings account deposits, and (**h**) NASDAQ Composite Index (*Source:* Adapted from "Statistics for Business and Economics." *TIME* Magazine, November 25, 1991. Copyright 1991 the Time, Inc. Magazine Company. Reprinted by permission)

Because interest rates such as that on 1-year CDs declined (specifically, from 8.67 % to 6.24 %) over the 3-year period, the proportion of savers' income derived from interest also went down (from 15 % in 1989 and 1990 to 14 % in 1991).

During these 3 years, time deposits and savings account deposits sank from $1,710 billion to $1,595 billion. Many savers withdrew their savings to invest them in the stock market. Consequently, the stock market indexes went up dramatically. For example, the NASDAQ Composite Index (an index compiled from over-the-counter stock) fluctuated from 454.8 in 1989 to 373.8 in 1990 to 531.3 in 1991.

As we can see in Fig. 2.8, bar charts can very effectively and clearly show changes in economic conditions. Common sense is all that the viewer needs to interpret the charts.

We discussed some of the data analysis related to Fig. 2.8a–h in this chapter; in later chapters, we will discuss it further, using more sophisticated statistical methods. These bar charts give us a great deal of information. And other statistical analysis related to this set of information will improve our understanding of macroeconomic analysis.

2.6 Summary

Good data are essential in business and economic decision making. Hence, it is important to be familiar with the sources of business and economic data and to know how these data can be collected.

Data for a census or a sample can be gleaned from both primary and secondary sources. However, we must guard against random error when using a sample and against systematic error in all our data collection.

Because we want to use sample data to make inferences about the population from which they are drawn, it is important for us to be able to present the data effectively. Tables and charts are two simple methods for presenting data. Line charts, bar charts, and pie charts are three basic and important graphical methods for describing data. In the next chapter, we will discuss other tabular and graphical methods for describing data in a more sophisticated and detailed manner.

Questions and Problems

1. What is a primary source of data? Give two examples of primary sources of data.
2. What is a secondary source of data? Give two examples of secondary sources of data.
3. What is a sample? What is a census? What advantages does using a sample have over using a census? Are there any advantages to using a census?
4. What two types of error might we encounter when dealing with primary and secondary sources of data?

5. Explain how the following can be used to present data.

   (a) Line chart
   (b) Component-parts line chart
   (c) Bar chart

6. Frederick Hallock is approaching retirement with a portfolio consisting of cash and money market fund investments worth $135,000, bonds worth $165,000, stocks worth $185,000, and real estate worth $1,200,000. Present these data in a bar chart.
7. LaPoint Glass Company has the following earnings before interest and taxes (EBIT) and profits (EBIT and profits are in millions of dollars).

Year	1988	1989	1990	1991
EBIT	3.3	3.3	4.1	5.5
Profits	1.6	1.8	2.1	2.8

   Present these data in a bar chart by hand and by using Lotus 1-2-3.
8. Of 354 MBA students, the following numbers chose to concentrate their study in these fields: 35 in finance, 63 in accounting, 70 in marketing, 35 in operations management, 52 in management information systems, 56 in economics, and 43 in organizational behavior. Present these data in a pie chart.
9. Use the data in Table 2.1 to draw line charts for the following:

   (a) GNP
   (b) CPI
   (c) GNP and CPI
   (d) 3-month T-bill rate and prime rate

10. Study Fig. 2.2, and comment on the relationship between GNP and private consumption.
11. Using the data in Fig. 2.3, analyze the average rates of return for

    (a) The DJIA
    (b) The S&P 500
    (c) Tri-Continental Corporation

12. Using the graph in Fig. 2.17, answer the following questions:

    (a) Which company has the higher current ratio?
    (b) Which company's current ratio appears to be more stable over time?

13. Using the graph in Fig. 2.19, carefully explain the relationship between Ford's inventory turnover and GM's.
14. You are given the following information about a certain company's current assets over the past 4 years:

	Years			
Current assets	1988	1989	1990	1991
Cash and marketable securities	4,215	5,341	6,325	5,842
Receivables	6,327	6,527	7,725	6,750
Inventories	9,254	9,104	10,104	11,100
Other current assets	2,153	3,277	4,331	3,956

Use a component-lines graph to plot this firm's current assets.

15. Explain under what conditions it is best to use a pie chart to present data.
16. Using the data given in question 14, present the components of total current assets for 1990 in two pie charts, one drawn by hand and one by using Microsoft Excel.
17. A statistics teacher has given the following numbers of the traditional grades to her class of 105 students:

Number of students	Grade
10	A
30	B
50	C
10	D
5	E

(a) Use a bar graph to show the distribution of grades.
(b) Use a pie chart to show the distribution of grades.
(c) Which of these graphs do you think is best for presenting the distribution of grades? Why?

18. Using the data in Table 2.5, show the distribution of current assets for 1996 in a pie chart and a bar chart. Which of these graphs do you think is best for presenting the data?
19. The following table gives the sales figures for five products manufactured by Trends Clothing Company, your employer.

Item	Sales
Sweaters	$5 million
Shirts	12 million
Pants	9 million
Blazers	16 million
Overcoats	7 million

The president of the company asks you for a report showing how sales are distributed among the five goods. What type of chart would you use?

20. Explain the benefits of graphs over tables in presenting data.
21. In the course of researching the benefits of diversification, you collect the information given in the table on page 47 (top), which presents rates of return for different portfolios.

(a) Use a line chart to plot the 20-year return for all five portfolios.
(b) What information do these plots provide?

Year	Stocks[a]	Bonds[b]	60 % stocks 40 % bonds	$\frac{1}{3}$ stocks $\frac{1}{3}$ bonds $\frac{1}{3}$ cash	BB&K Index[c]
1970	4.01 %	12.10 %	7.52 %	7.98 %	4.7 %
1971	14.31	13.23	14.14	10.83	13.7
1972	18.98	5.68	13.54	9.38	15.1
1973	−14.66	−1.11	−9.11	−3.03	−2.2
1974	−26.47	4.35	−14.88	−5.44	−6.6
1975	37.20	9.19	25.65	17.04	19.6
1976	23.84	16.75	21.18	15.19	11.5
1977	−7.18	−.67	−4.57	−0.94	6.1
1978	6.56	−1.16	3.65	4.40	13.0
1979	18.44	−1.22	10.28	9.14	11.5
1980	32.42	−3.95	17.45	13.17	17.9
1981	−4.91	1.85	−1.99	4.06	6.4
1982	21.41	40.35	28.98	23.97	14.4
1983	22.51	.68	13.43	10.52	15.4
1984	6.27	15.43	10.11	10.75	10.4
1985	32.16	30.97	31.85	23.38	25.4
1986	18.47	24.44	21.11	16.61	23.3
1987	5.23	−2.69	3.59	3.92	8.6
1988	16.81	9.67	13.97	11.01	13.2
1989	31.49	18.11	26.24	19.22	14.3
Compound annual return	11.55	9.00	10.89	9.78	11.54

Source: Bailard, Biehl & Kaiser, Ibbotson Associates. Inc. This figure was printed in the *Wall Street Journal* on January 25, 1990, p. Cl
[a]Standard & Poor's 500 index
[b]Long-term Treasury bonds
[c]20 % US stocks, 20 % bonds, 20 % cash, 20 % real estate, 20 % foreign stocks

22. Use the data given in question 21 to construct a bar graph for 1985 through 1989.
23. You are given the following exchange rate information for the number of dollars it takes to buy 1 British pound and the number of dollars it takes to buy 100 Japanese yen:

Month	$/BP	$/100 yen
Jan 88	1.7505	.7722
Feb 88	1.7718	.7782
Mar 88	1.8780	.8042
Apr 88	1.8825	.8015
May 88	1.8410	.7995
Jun 88	1.7042	.7475
Jul 88	1.7160	.7533

(continued)

Month	$/BP	$/100 yen
Aug 88	1.6808	.7307
Sep 88	1.6930	.7477
Oct 88	1.7670	.7951
Nov 88	1.8505	.8227
Dec 88	1.8075	.8013

(a) Draw a line chart showing the exchange rates between British pounds (BP) and US dollars during this period.
(b) Draw a line chart showing the exchange rates between Japanese yen and US dollars.
(c) Use Microsoft Excel to draw a line chart containing the exchange rates in *a* and a line chart representing the exchange rates in *b*.

24. You are given the following financial ratios for Johnson & Johnson and for the pharmaceutical industry:

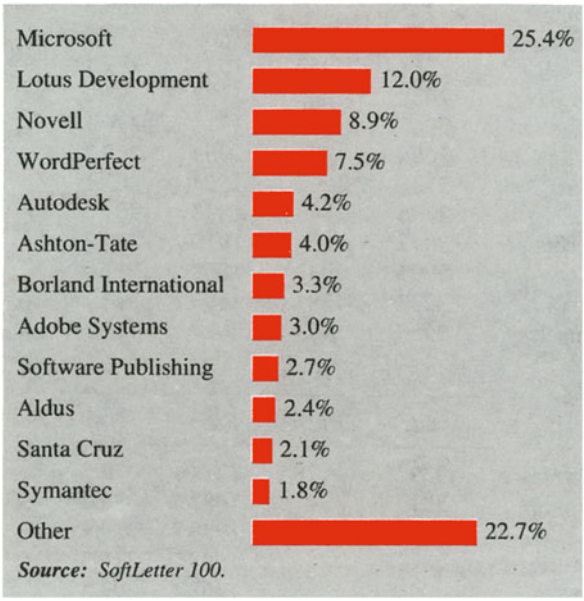
	Current ratio		*Inventory Turnover*	
Year	Industry	J&J	Industry	J&J
79	2.30	2.73	2.17	2.71
80	2.29	2.55	2.22	2.70
81	2.18	2.50	2.34	2.78
82	2.12	2.50	2.30	2.38
83	2.12	2.66	2.34	2.28
84	2.09	2.41	2.40	2.37
85	2.19	2.47	2.27	2.45
86	1.91	1.40	2.38	2.33
87	1.86	1.86	2.24	2.27
88	1.93	1.88	2.27	2.32

	ROA		*Price/earnings*	
Year	Industry	J&J	Industry	J&J
79	.11	.12	12.04	13.76
80	.11	.12	15.03	15.35
81	.10	.12	28.02	14.79
82	.11	.12	15.19	17.79
83	.11	.11	14.56	15.90
84	.10	.11	14.88	13.14
85	.10	.12	17.98	15.66
86	.10	.06	23.46	35.47
87	.09	.13	35.09	15.50
88	.12	.14	15.98	14.88

(a) Draw a line chart showing the current ratio over time for the industry and for J&J, and compare the two.
(b) Use a bar graph to present the data for the industry and J&J's current ratio.

25. Repeat question 24 for inventory turnover.
26. Repeat question 24 for return on total assets (ROA).
27. Repeat question 24 for the price/earnings ratio.
28. An August 27, 1991, *Wall Street Journal* article reported that increasing numbers of small software firms are being absorbed by that industry's biggest companies. According to WSJ, the result of this dominance by a few giants is that the industry has become much tougher for software entrepreneurs to break into. The newspaper printed the chart given in the accompanying figure to depict the breakdown of market share among software companies. Refer to this chart to answer the following questions:

    (a) List the companies in descending order of market share.
    (b) What is the combined market share for Lotus Development and WordPerfect?
    (c) What is the combined market share for Micro soft, Lotus Development, and Novell?



    From entrepreneurs to corporate giants: market share among the top 100 software companies, based on total 1990 revenue of $5.7 billion.

29. The results of the 1991 city council election (voters could vote for more than one person) in Monroe Township, New Jersey, were

Nalitt	4,656
Riggs	4,567
Anderson	4,140
Miller-Paul	4,142
	17,505

Use a pie chart to present the results of the election.

30. Redo question 29 using a bar chart. Which method is better for presenting these election results?

To answer questions 31–37, refer to the table, which gives the rankings for team defense and offense for NFC teams for the first 9 weeks of the 1991 season.

Rankings of team defense and offense for NFC teams in the 1991 season (rankings based on averages a game)

	NFC team defense			
	Yds	*Rush*	*Pass*	*Avg.*
Philadelphia	1,955	715	1,240	217.2
New Orleans	2,035	562	1,473	226.1
Washington	2,325	830	1,495	258.3
San Francisco	2,460	851	1,609	273.3
New York	2,551	959	1,592	283.4
Tampa Bay	2,652	1,065	1,587	294.7
Chicago	2,665	950	1,715	296.1
Green Bay	2,684	775	1,909	298.2
Atlanta	2,728	1,202	1,526	303.1
Dallas	2,736	863	1,873	304.0
Minnesota	3,097	1,147	1,950	309.7
Detroit	2,799	932	1,867	311.0
Phoenix	3,277	1,381	1,896	327.7
Los Angeles	2,986	959	2,027	331.8

	NFC team offense			
	Yds	*Rush*	*Pass*	*Avg.*
San Francisco	3,392	1,178	2,214	376.9
Washington	3,019	1,337	1,682	335.4
Dallas	2,969	970	1,999	329.9
New York	2,842	1,254	1,588	315.8
Minnesota	3,095	1,328	1,767	309.5
Atlanta	2,768	998	1,770	307.6
Detroit	2,705	1,070	1,635	300.6
New Orleans	2,665	918	1,747	296.1
Chicago	2,662	1,006	1,656	295.8
Los Angeles	2,515	748	1,767	279.4
Phoenix	2,636	897	1,739	263.6
Philadelphia	2,319	688	1,631	257.7
Green Bay	2,250	650	1,600	250.0
Tampa Bay	2,142	779	1,363	238.0

Source: USA TODAY, November 7, 1991, p. 11C

31. Use a pie chart to show how San Francisco's total team offense is divided between rush and pass.
32. Use a pie chart to show how Phoenix's total team defense is divided between rush and pass.
33. Use a bar chart to show the total pass offense for the 14 NFC teams.
34. Repeat question 33 for rush offense.
35. Repeat question 33 for pass defense.
36. Repeat question 33 for rush defense.
37. Use the graphs from questions 33–36 to answer the following questions;

    (a) Which team has the best pass offense?
    (b) Which team has the best pass defense?
    (c) Which team has the best rush offense?
    (d) Which team has the best rush defense?

38. The following table is a table of salaries for the top NHL forwards and defensemen:

    (a) Use a bar chart to show the players' salaries.
    (b) Do you think the bar chart is a better vehicle than a table for comparing players' salaries?

    Salary comparisons for top NHL forwards and defensemen

Position	Name	Team	Gross salary ($ millions)
C	Wayne Gretzky	Los Angeles Kings	$3
C	Mario Lemieux	Pittsburgh Penguins	$2.338
RW	Brett Hull	St. Louis Blues	$1.5
C	Pat LaFontaine	Buffalo Sabres	$1.4
C	Steve Yzerman	Detroit Red Wings	$1.4
LW	Kevin Stevens	Pittsburgh Penguins	$1.4
LW	Luc Robitaille	Los Angeles Kings	$1.3
C	John Cullen	Hartford Whalers	$1.2
D	Ray Bourque	Boston Bruins	$1.2
D	Scott Stevens	New Jersey Devils	$1.155

Source: USA TODAY, October 7, 1991, p. 8C

39. The accompanying pie chart presents data on why teenagers drink. Use information shown in the pie chart to answer the following questions:

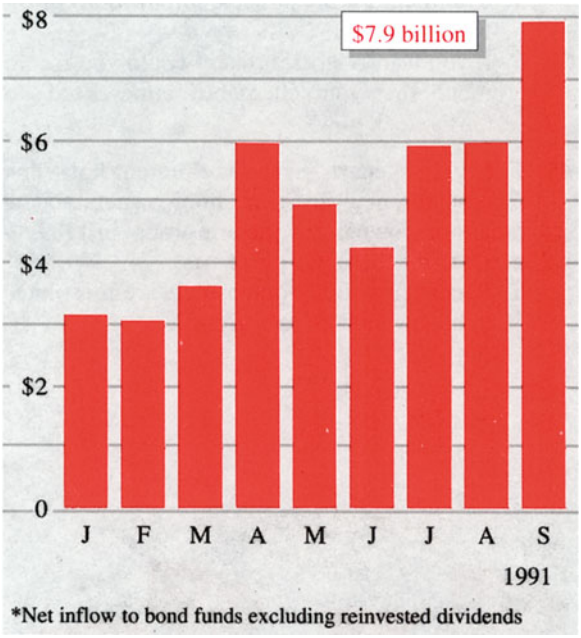
    (a) For what reason do the highest numbers of teenagers drink?
    (b) What percentage of teenagers drink because they are bored or because they are upset?

Source: National Council on Alcoholism and Drug Dependence. Surgeon General survey. *USA TODAY*, November 5, 1991. Copyright 1991, USA TODAY. Reprinted with permission.

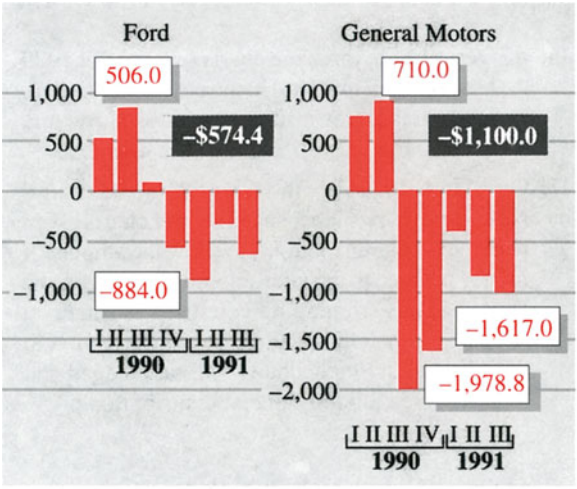
To answer questions 40–42, use the following results of the election to the General Assembly from one New Jersey district in 1991.

Batten	17,026
Lookabaugh	17,703
LoBiondo	27,452
Gibson	24,735

40. Use a bar graph to show the distribution of votes.
41. Use a pie chart to show the distribution of votes.
42. Which type of graph presents these data more effectively?
43. The following bar graph shows net purchases of bond mutual funds. Does this graph tell us anything?



44. On November 9, 1991, the *Home News* of central New Jersey used the bar chart given in the accompanying figure to show quarterly net income or losses for both Ford and GM.

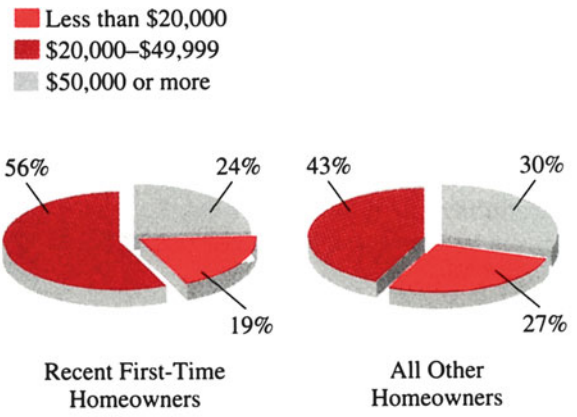


Source: Company reports, news reports. Reprinted by permission of *Knight-Ridder Tribune News.*

(a) Comment on the possible implications of this bar chart.
(b) If you were a stock broker, would you recommend that your client buy either Ford's or GM's stock now?

45. The two pie charts given here present household income for new first-time homeowners and all other homeowners, by income group, in 1989.

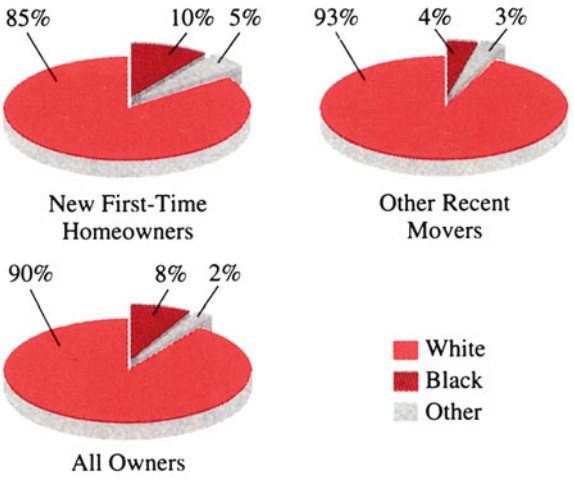
(a) Describe these two pie charts.
(b) Recent first-time homeowners are most likely to be in which income group?



Source: U.S. Census Bureau. AP. *Home News*, October 31, 1991.

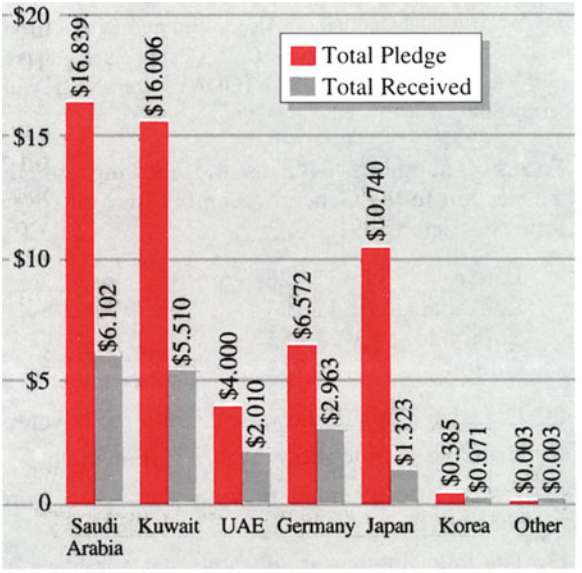
46. (a) Describe the three pie charts (in terms of 1989 home ownership data) in the next column (top).
(b) Do members of minority groups show any gains among new first-time homeowners?

47. On March 20, 1991, the *Home News* (a central New Jersey newspaper) used the bar charts given here (next column, bottom) to show the amount of money pledged to, and the amount received by, the United States from allied countries as financial support for the Gulf War. Use the information in this chart to draw pie charts of total pledged and total received allied financial contribution.
Analysis. (Hint: Use the pie chart.)



Source: U.S. Census Bureau. *Home News,* October 31, 1991. AP/Ed De Gasero, reprinted by permission of The Associated Press.

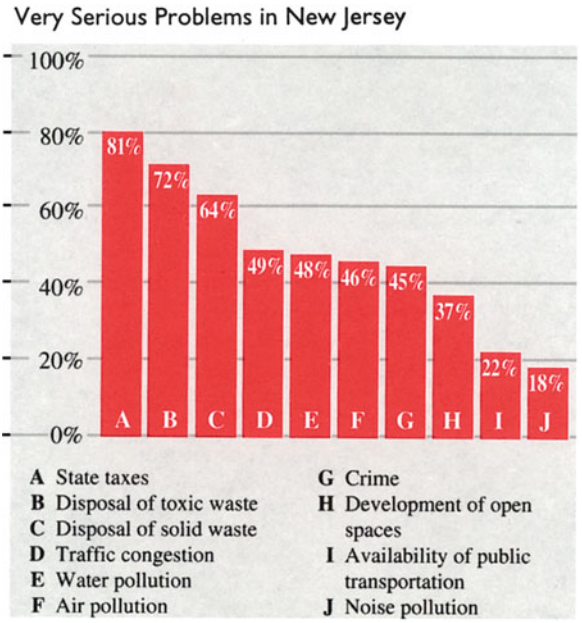
Figures in billions of dollars



Source: Senate Appropriations Committee. *Home News,* March 20, 1991. Reprinted by permission of The Associated Press.

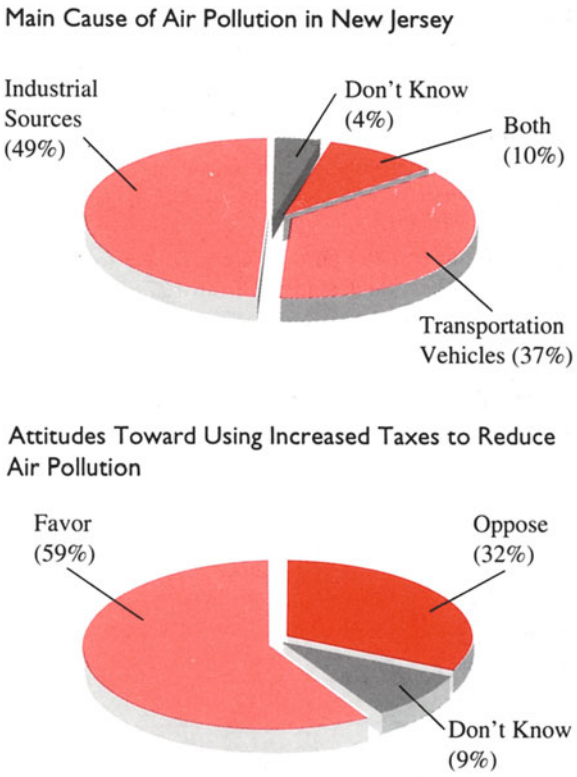
48. On March 14, 1991, the *Home News* (a central New Jersey newspaper) used the bar chart given here to show what problems New Jerseyans considered "very serious."

    (a) What do New Jerseyans consider the most serious problem?
    (b) Is traffic congestion regarded as more serious than crime? Explain your answer.



Source: Project: CLEAN AIR/Eagleton Institute of Politics. *Home News,* March 14, 1991. Reprinted with permission of the publisher.

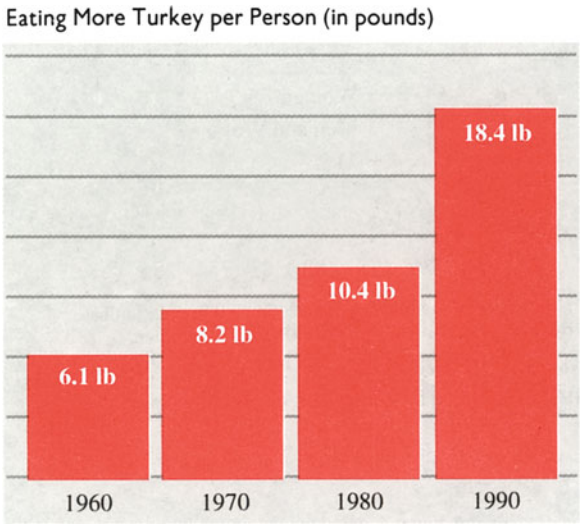
49. On March 14, 1991, the *Home News* used two pie charts (next column, top) to show (1) the main causes of air pollution in New Jersey and (2) people's attitudes toward using increased taxes to reduce air pollution. Discuss the implications of these two pie charts and of the bar chart given in question 48.



Source: Project: CLEAN AIR/Eagleton Institute of Politics. From *Home News,* March 14, 1991. Reprinted with permission of the publisher.

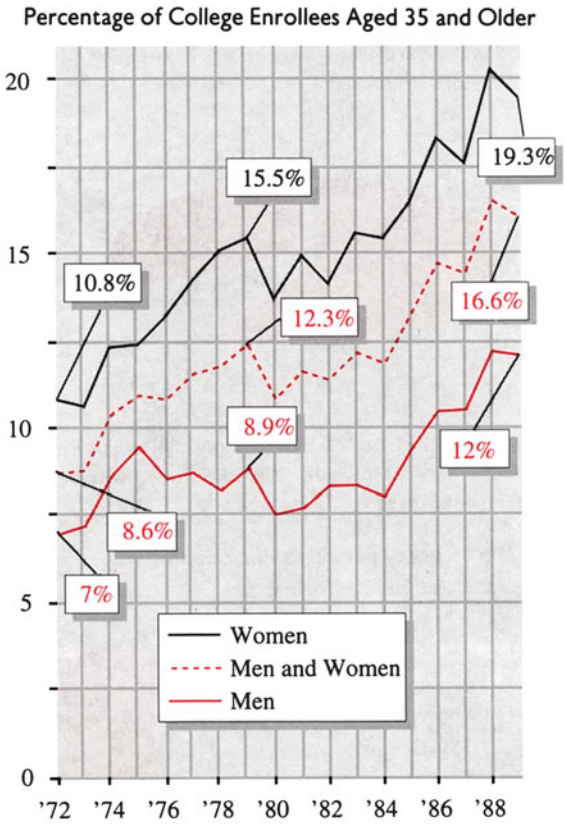
50. The *Home News* used this bar chart (next column, bottom) on page Dl of its November 20, 1991, issue to depict the increasing popularity of turkey not just at holiday meals but throughout the year.

    (a) How much turkey was consumed per person in1960–1990, respectively?
    (b) How much has per person consumption of turkey increased from 1970 to 1990?



51. The following line chart was printed in the *Home News* on page Al of its November 22, 1991, issue to show the increase in the number of college students over age 35 during the period of 1972–1989.

    (a) Percentage wise, did more men or more women over 35 years old attend college during this 18-year period?
    (b) What was the percentage increase for older female college students from 1977 to 1989?
    (c) What was the percentage increase for college students over age 35 from 1972 to 1989?



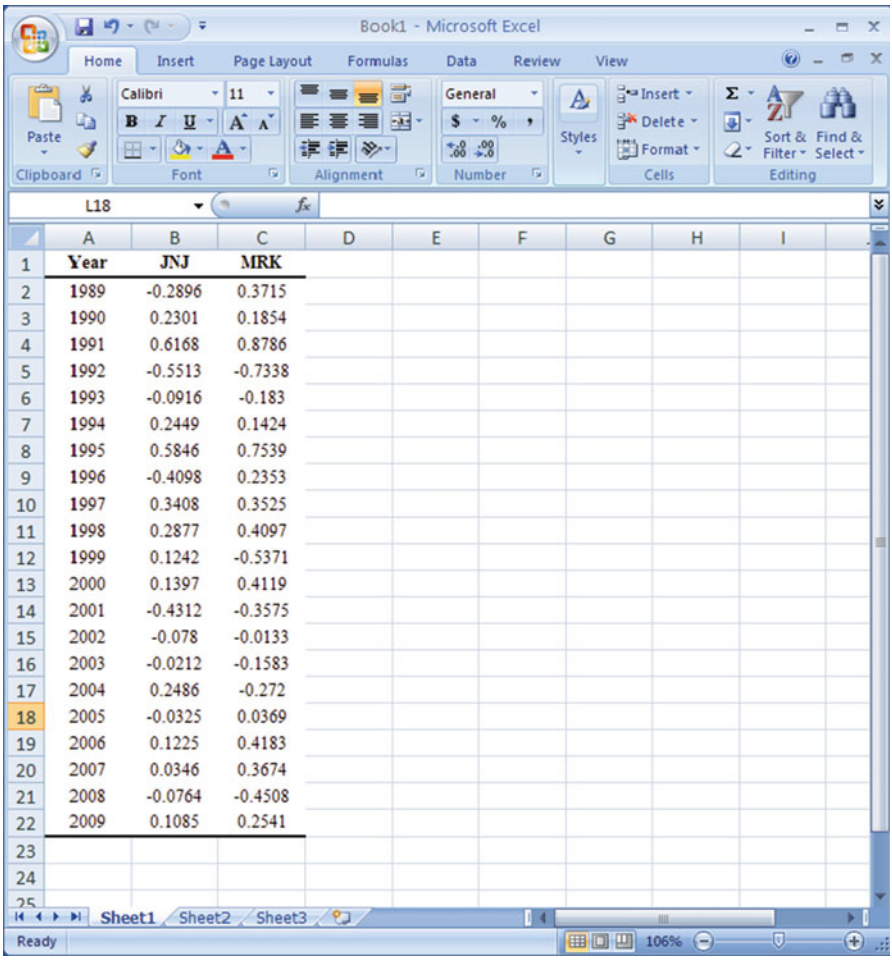
Note: Several methods used on estimated population base by the U.S. Census Bureau to calculate percentage.

Source: U.S. Census Bureau. *Home News,* November 22, 1991. Reprinted by permission of The Associated Press.

Appendix 1: Using Microsoft Excel to Draw Graphs

This appendix explains how to use Microsoft Excel to draw graphs. Eight steps are involved: entering Microsoft Excel, preparing the data, and drawing the graph(s):

Stage 1: Start Excel and enter data for Johnson & Johnson and Merck as shown in Fig. 2.9.
Stage 2: Select the data to be graphed as shown in Fig. 2.10.
Stage 3: From the Insert Menu, choose Line on charts option as shown in Fig. 2.11 to get the 2-D line chart as shown in Fig. 2.12.
Stage 4: Choose the Select Data on Chart Tools as shown in Fig. 2.12.



	A	B	C
1	Year	JNJ	MRK
2	1989	-0.2896	0.3715
3	1990	0.2301	0.1854
4	1991	0.6168	0.8786
5	1992	-0.5513	-0.7338
6	1993	-0.0916	-0.183
7	1994	0.2449	0.1424
8	1995	0.5846	0.7539
9	1996	-0.4098	0.2353
10	1997	0.3408	0.3525
11	1998	0.2877	0.4097
12	1999	0.1242	-0.5371
13	2000	0.1397	0.4119
14	2001	-0.4312	-0.3575
15	2002	-0.078	-0.0133
16	2003	-0.0212	-0.1583
17	2004	0.2486	-0.272
18	2005	-0.0325	0.0369
19	2006	0.1225	0.4183
20	2007	0.0346	0.3674
21	2008	-0.0764	-0.4508
22	2009	0.1085	0.2541

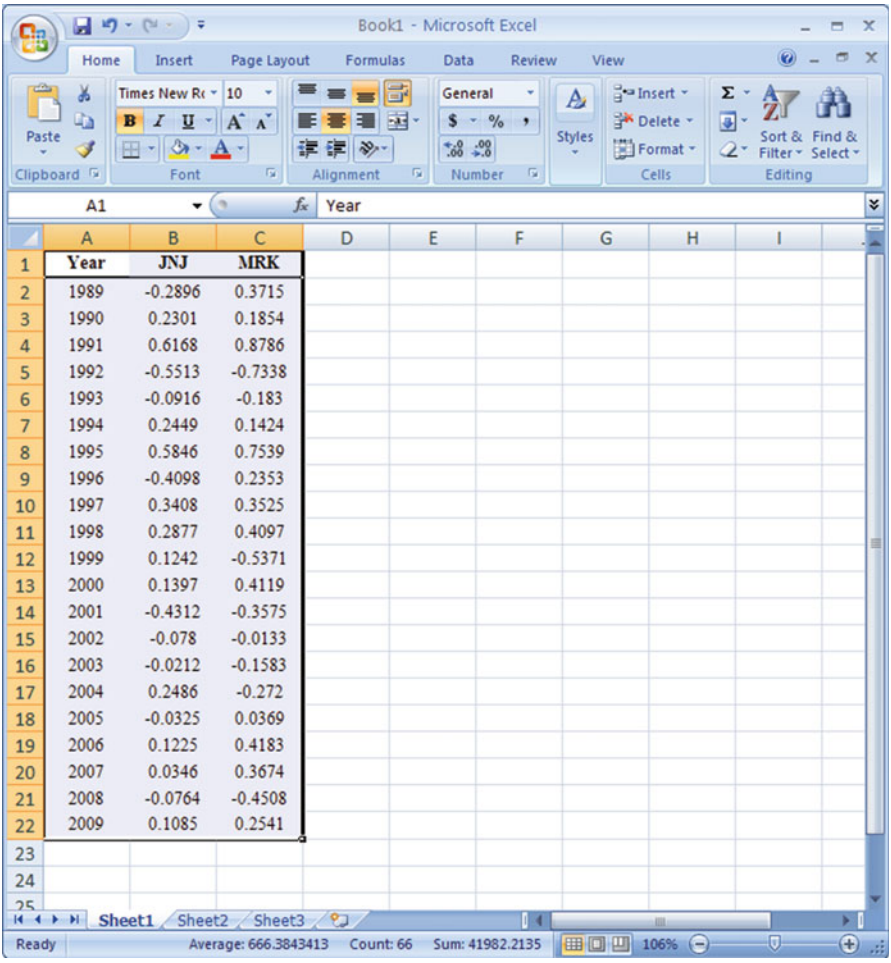
Fig. 2.9 Rates of return for JNJ and MRK

Stage 5: Choose Year and press Remove as shown in Fig. 2.13. Then press Edit to select axis labels as shown in Fig. 2.14.

Stage 6: Select Cell A2 to Cell A22 in Axis label range as shown in Fig. 2.14 and press OK.

Stage 7: Press OK on Select Data Source again, and then we can get the chart as shown in Fig. 2.15. Press Move Chart on Chart Tools, select New sheet, and press OK. The finished chart is shown in Fig. 2.16.

Microsoft Excel has strong charting features. With more work, the chart in Fig. 2.16 can look like the chart in Fig. 2.17.



	A	B	C
1	Year	JNJ	MRK
2	1989	-0.2896	0.3715
3	1990	0.2301	0.1854
4	1991	0.6168	0.8786
5	1992	-0.5513	-0.7338
6	1993	-0.0916	-0.183
7	1994	0.2449	0.1424
8	1995	0.5846	0.7539
9	1996	-0.4098	0.2353
10	1997	0.3408	0.3525
11	1998	0.2877	0.4097
12	1999	0.1242	-0.5371
13	2000	0.1397	0.4119
14	2001	-0.4312	-0.3575
15	2002	-0.078	-0.0133
16	2003	-0.0212	-0.1583
17	2004	0.2486	-0.272
18	2005	-0.0325	0.0369
19	2006	0.1225	0.4183
20	2007	0.0346	0.3674
21	2008	-0.0764	-0.4508
22	2009	0.1085	0.2541

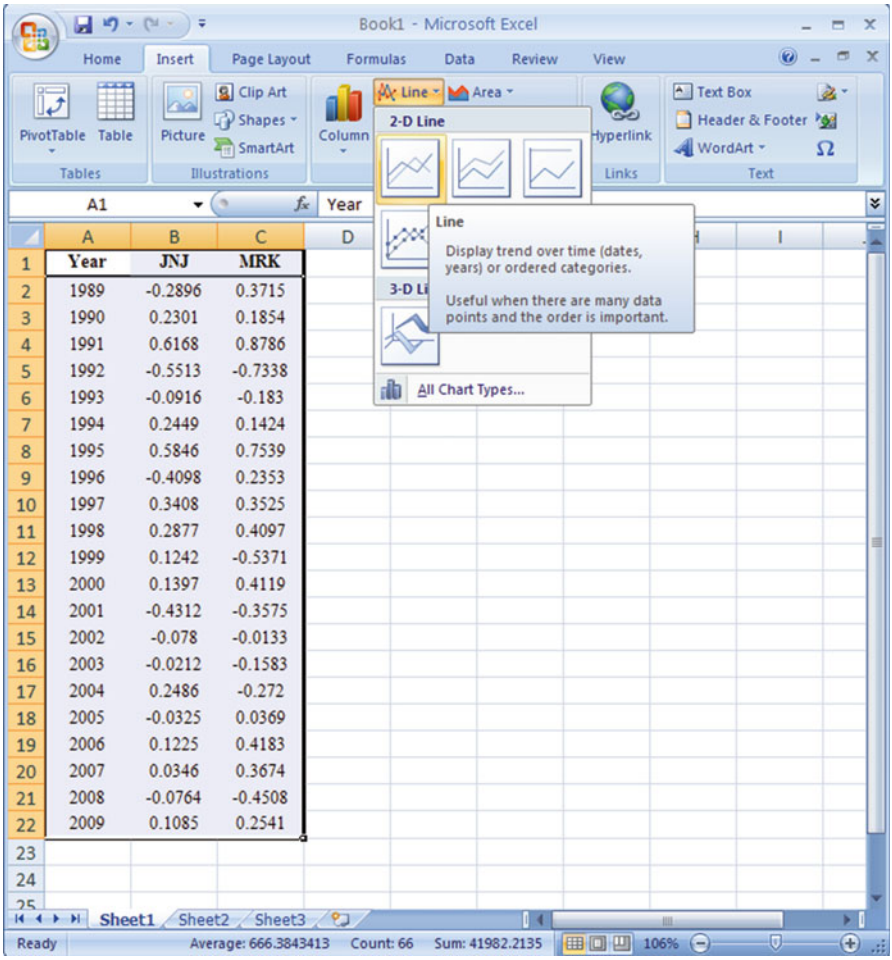
Fig. 2.10 Data to be graphed

Appendix 2: Stock Rates of Return and Market Rates of Return

Table 2.3 presents data on earnings per share (EPS), dividends per share (DPS), and price per share (PPS) for Johnson & Johnson, Merck, and the S&P 500 during the period 1988–2009. Table 2.4 shows rates of return for Johnson & Johnson, Merck, and the S&P 500, calculated from the data in Table 2.3.

The formula for calculating the rate of return, R_{jt}, on the jth individual stock in period t is

$$R_{jt} = \frac{P_{jt} - P_{jt-1} + d_{jt}}{P_{jt-1}} \tag{2.1}$$



	A	B	C
1	Year	JNJ	MRK
2	1989	-0.2896	0.3715
3	1990	0.2301	0.1854
4	1991	0.6168	0.8786
5	1992	-0.5513	-0.7338
6	1993	-0.0916	-0.183
7	1994	0.2449	0.1424
8	1995	0.5846	0.7539
9	1996	-0.4098	0.2353
10	1997	0.3408	0.3525
11	1998	0.2877	0.4097
12	1999	0.1242	-0.5371
13	2000	0.1397	0.4119
14	2001	-0.4312	-0.3575
15	2002	-0.078	-0.0133
16	2003	-0.0212	-0.1583
17	2004	0.2486	-0.272
18	2005	-0.0325	0.0369
19	2006	0.1225	0.4183
20	2007	0.0346	0.3674
21	2008	-0.0764	-0.4508
22	2009	0.1085	0.2541

Fig. 2.11 Select graph option

where P_{jt} represents price per share for the jth stock in period t and d_{jt} represents dividends per share for the jth stock. The market rate of return, R_{mt}, in period t is

$$\frac{\mathrm{SP}_t - \mathrm{SP}_{t-1}}{\mathrm{SP}_{t-1}} \tag{2.2}$$

where SP_t represents the S&P 500 in period t.

The rate of return on an individual stock can be rewritten as

$$R_{jt} = \frac{(P_{jt} - P_{jt+1})}{P_{jt-1}} + \frac{d_{jt}}{P_{jt-1}} = \text{Capital gain yield} + \text{dividend yield} \tag{2.3}$$

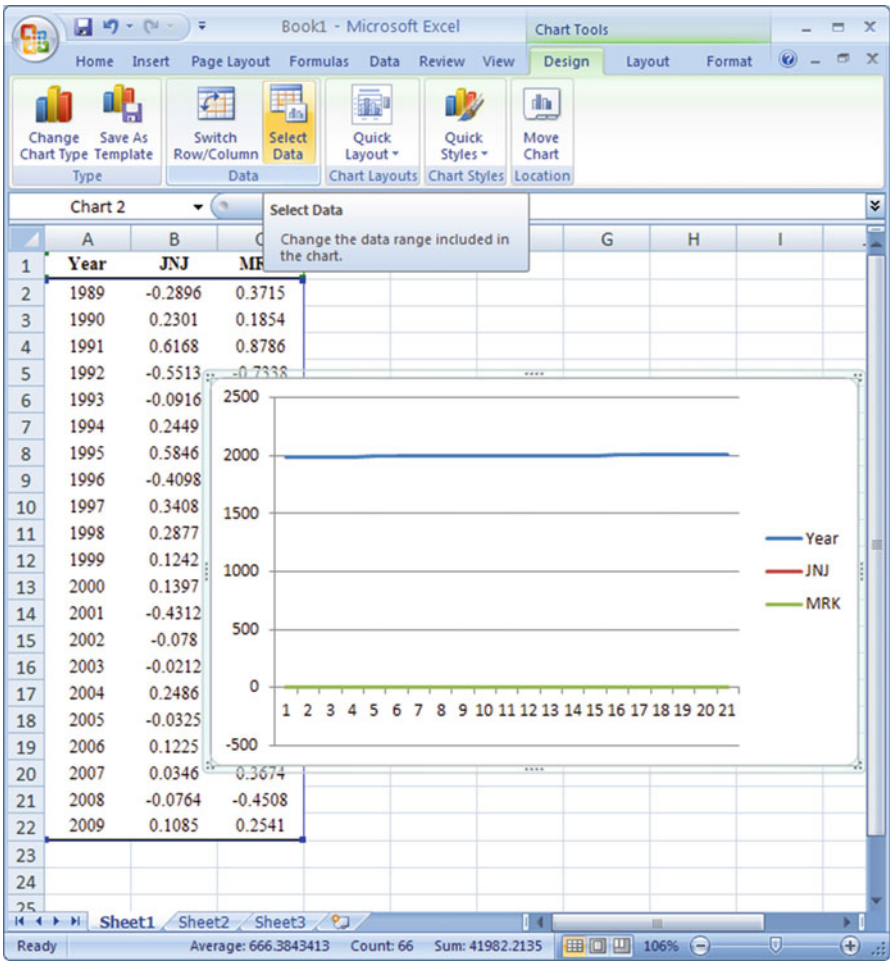


Fig. 2.12 Step 1 of Chart Tools

The first term of the rewritten equation is the capital gain yield (in percent); the second is the dividend yield (also in percent). As an example, let us calculate the rate of return for Johnson & Johnson in 2009. From Table 2.3, we know that $PPS_{08} = \$59.83$, $PPS_{09} = \$64.41$, and $DPS_{09} = \$1.91$. Thus, the rate of return for Johnson & Johnson in 2009 equals

$$R_{JNJ09} = \frac{64.41 - 59.83 + 1.91}{59.83} = .1085.$$

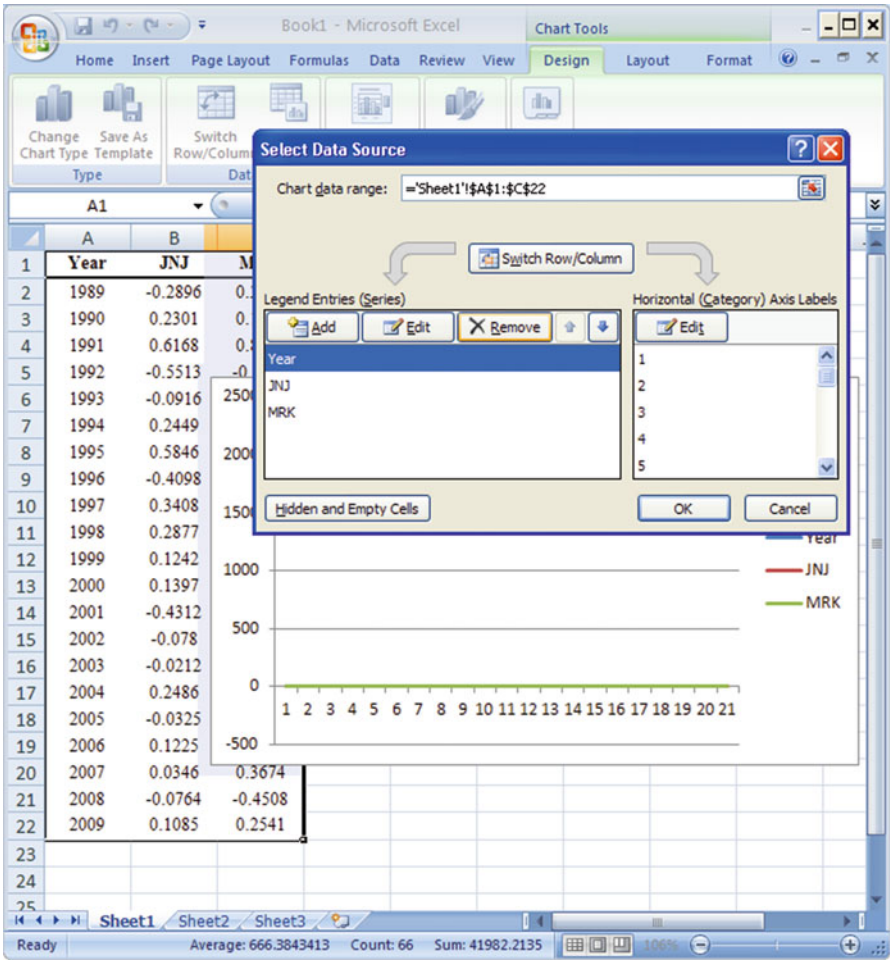
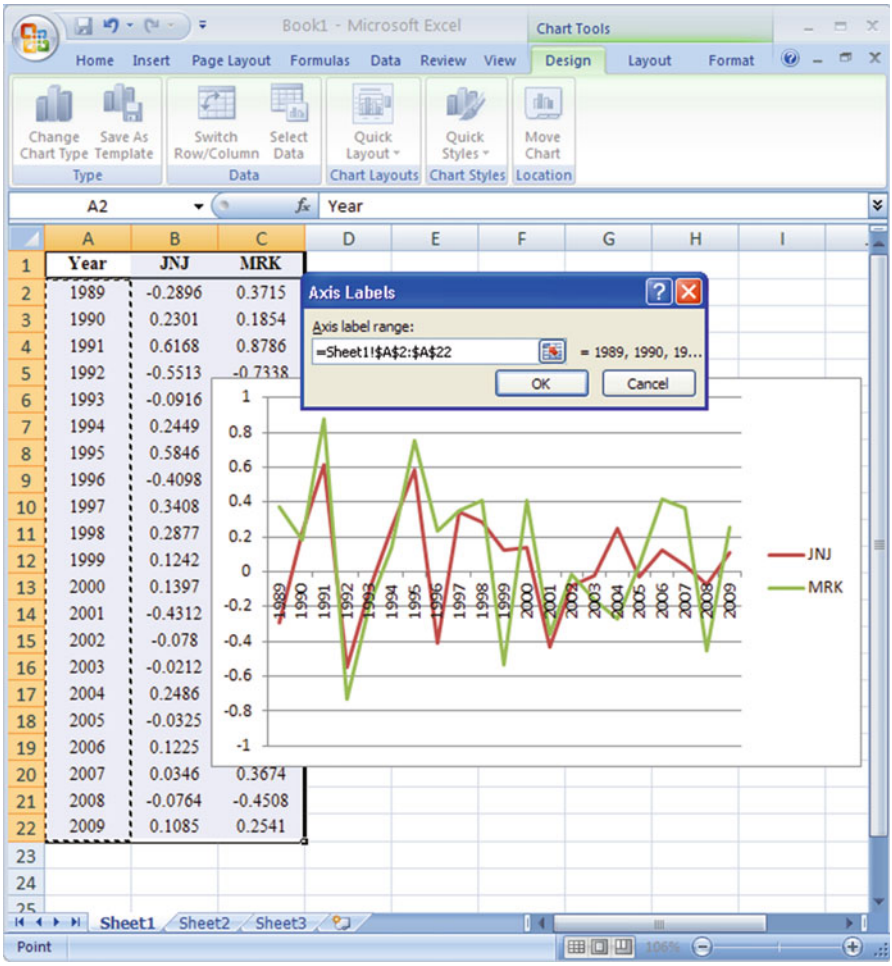


Fig. 2.13 Step 2 of Chart Tools

As another example, from Table 2.3, we know that the S&P 500 was 1220.04 and 948.05 in 2008 and 2009, respectively. Thus, the market rate of return in 2008 equaled

$$R_{M08} = \frac{948.05 - 1220.04}{1220.04} = -0.2229$$

Figure 2.5 in the text compares the rates of return for Johnson & Johnson, Merck, and the S&P 500 over time, as discussed in Sect. 2.5 of this chapter.



	A	B	C
1	Year	JNJ	MRK
2	1989	-0.2896	0.3715
3	1990	0.2301	0.1854
4	1991	0.6168	0.8786
5	1992	-0.5513	-0.7338
6	1993	-0.0916	
7	1994	0.2449	
8	1995	0.5846	
9	1996	-0.4098	
10	1997	0.3408	
11	1998	0.2877	
12	1999	0.1242	
13	2000	0.1397	
14	2001	-0.4312	
15	2002	-0.078	
16	2003	-0.0212	
17	2004	0.2486	
18	2005	-0.0325	
19	2006	0.1225	
20	2007	0.0346	0.3674
21	2008	-0.0764	-0.4508
22	2009	0.1085	0.2541

Fig. 2.14 Step 3 of Chart Tools

Appendix 3: Financial Statements and Financial Ratio Analysis

Review of Balance Sheets and Income Statements

Accounting concepts are used to understand a firm's financial condition. We will discuss two basic sources of accounting information: the *balance sheet*, which reveals the assets, liabilities, and owners' (stockholders') equity of a firm *at* a *point* in time, and the income statement, which shows the firm's profit or loss *over a* given *period* of time. *Assets*, which are things that the firm owns, can be classified as current, fixed, or "other" assets. Current assets consist of cash and

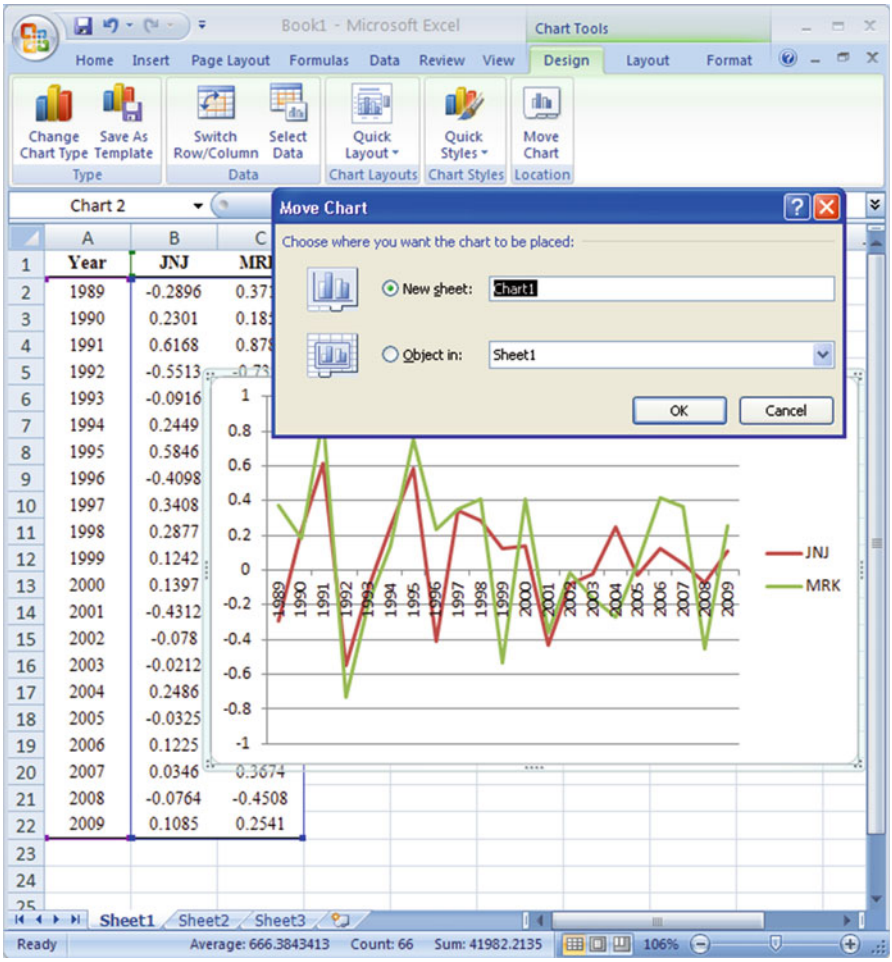


Fig. 2.15 Step 4 of Chart Tools

of property that can be turned into cash quickly. Examples of current assets include cash, stocks, bonds, inventory, and accounts receivable (cash that customers owe the firm). Fixed assets are not easily convertible into cash; they include land, machinery, and buildings. Fixed assets are generally valued at their historical value (purchase price) minus depreciation, not at their current market value. Other assets include intangibles such as goodwill, trademarks, patents, copyrights, and leases.

Liabilities, which are debts of the firm, are divided into current and long-term debts. Current debts come due within 1 year, whereas long-term debts are due in more than 1 year. Current debts include accounts payable (unpaid bills), notes payable, debts on agreements, accrued expenses, expenses incurred but not yet

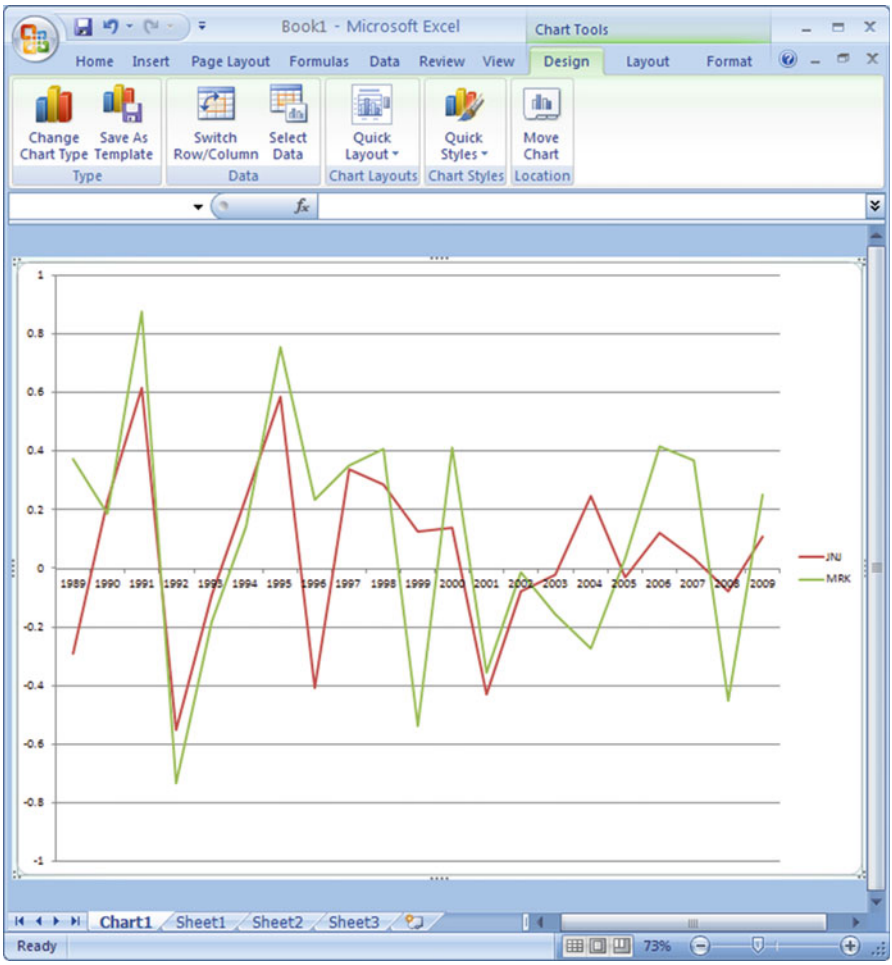


Fig. 2.16 Line Charts of rates of return for JNJ and MRK

paid, and taxes payable. Examples of long-term liabilities include mortgages, payable corporate bonds, and capitalized leases.

Stockholders' equity makes up the second part of the liabilities section of a balance sheet. It consists of funds invested by shareholders plus retained earnings. The *net worth* of the firm is calculated by subtracting total liabilities from total assets.

Whereas a balance sheet looks at the firm at a point in time, the income statement, as we noted earlier, evaluates the firm over a period of time. The end product of the income statement is the profit or loss, which is calculated by taking the sales for a period and subtracting the cost of goods sold and such expenses as

Table 2.3 EPS, DPS, and PPS for Johnson & Johnson, Merck, and the S&P 500

	Johnson & Johnson			Merck			
Year	*DPS*	*EPS*	*PPS*	*DPS*	*EPS*	*PPS*	S&P500
1988	1.89	5.63	85.13	1.37	3.02	57.75	265.79
1989	1.10	3.19	59.38	1.70	3.74	77.50	322.84
1990	1.29	3.38	71.75	2.00	4.51	89.88	334.59
1991	1.51	4.30	114.50	2.34	5.39	166.50	376.18
1992	0.88	1.54	50.50	0.95	1.70	43.38	415.74
1993	1.00	2.71	44.88	1.06	1.86	34.38	451.41
1994	1.12	3.08	54.75	1.15	2.35	38.13	460.42
1995	1.25	3.65	85.50	1.24	2.63	65.63	541.72
1996	0.72	2.12	49.75	1.44	3.12	79.63	670.5
1997	0.83	2.41	65.88	1.70	3.74	106.00	873.43
1998	0.95	2.23	83.88	1.93	4.30	147.50	1,085.5
1999	1.04	2.94	93.25	1.09	2.45	67.19	1,327.33
2000	1.22	3.39	105.06	1.23	2.90	93.63	1,427.22
2001	0.66	1.83	59.10	1.36	3.14	58.80	1,194.18
2002	0.78	2.16	53.71	1.41	3.14	56.61	993.94
2003	0.91	2.39	51.66	1.45	3.03	46.20	965.23
2004	1.08	2.83	63.42	1.50	2.61	32.14	1,130.65
2005	1.26	3.46	60.10	1.52	2.10	31.81	1,207.23
2006	1.44	3.73	66.02	1.52	2.03	43.60	1,310.46
2007	1.60	3.63	66.70	1.51	1.49	58.11	1,477.19
2008	1.77	4.57	59.83	1.52	3.64	30.40	1,220.04
2009	1.91	4.40	64.41	1.58	5.68	36.54	948.05

Source: EPS, DPS, and PPS for Johnson & Johnson and Merck are from Standard & Poor's Compustat, Wharton Research Data Services (WRDS)

Table 2.4 Rates of return for Johnson & Johnson and Merck stock and the S&P 500

Year	JNJ	MRK	S&P500
1989	−0.2896	0.3715	0.2146
1990	0.2301	0.1854	0.0364
1991	0.6168	0.8786	0.1243
1992	−0.5513	−0.7338	0.1052
1993	−0.0916	−0.1830	0.0858
1994	0.2449	0.1424	0.0200
1995	0.5846	0.7539	0.1766
1996	−0.4098	0.2353	0.2377
1997	0.3408	0.3525	0.3027
1998	0.2877	0.4097	0.2428
1999	0.1242	−0.5371	0.2228
2000	0.1397	0.4119	0.0753
2001	−0.4312	−0.3575	−0.1633
2002	−0.0780	−0.0133	−0.1677
2003	−0.0212	−0.1583	−0.0289
2004	0.2486	−0.2720	0.1714
2005	−0.0325	0.0369	0.0677
2006	0.1225	0.4183	0.0855
2007	0.0346	0.3674	0.1272
2008	−0.0764	−0.4508	−0.1741
2009	0.1085	0.2541	−0.2229

Table 2.5 Johnson & Johnson corporation balance sheet ($ million)

Assets	2004	2005	2006	2007	2008	2009
Current assets						
Cash and cash equivalent	$9,203	$16,055	$4,083	$7,770	$10,768	$15,810
Marketable securities	3,681	83	1	1,545	2,041	3,615
Account receivable	6,831	7,010	8,712	9,444	9,719	9,646
Inventory	3,744	3,959	4,889	5,110	5,052	5,180
Deferred taxes on income	1,737	1,845	2,094	2,609	3,430	2,793
Prepaid expenses and other receivable	2,124	2,442	3,196	3,467	3,367	2,497
Total current assets	*27,320*	*31,394*	*22,975*	*29,945*	*34,377*	*39,541*
Marketable securities—noncurrent	46	20	16	2	–	–
Property, plant and equipment, net	10,436	10,830	13,044	14,185	14,365	14,759
Intangible assets, net	11,842	12,175	28,688	28,763	27,695	31,185
Deferred taxes on income	551	385	3,210	4,889	5,841	5,507
Other assets	3,122	3,221	2,623	3,170	2,634	3,690
Total assets	*53,317*	*58,025*	*70,556*	*80,954*	*84,912*	*94,682*
Liabilities and shareholder's equity						
Current liabilities						
Loans and notes payable	280	668	4,579	2,463	3,732	6,318
Account payable	5,227	4,315	5,691	6,909	7,503	5,541
Accrued liabilities	3,523	3,529	4,587	6,412	5,531	5,796
Accrued rebates, returns, and promotion	2,297	2,017	2,189	2,318	2,237	2,028
Accrued salaries, wages, and commissions	1,094	1,166	1,391	1,512	1,432	1,606
Taxes on income	1,506	940	724	223	417	442
Total current liabilities	*13,927*	*12,635*	*19,161*	*19,837*	*20,852*	*21,731*
Long-term debt	2,565	2,017	2,014	7,074	8,120	8,223
Deferred tax liability	403	211	1,319	1,493	1,432	1,424
Employee-related obligations	2,631	3,065	5,584	5,402	7,791	6,769
Other liabilities	1,978	2,226	3,160	3,829	4,206	5,947
Shareowners' equity						
Preferred stock—without par value	–	–	–	–	–	–
Common stock–par value $1.00	3,120	3,120	3,120	3,120	3,120	3,120
Net receivable from employee stock plan	−11	–	–	–	–	–
Accumulated other comprehensive income	−515	−755	−2,118	−693	−4,955	−3,058
Retained earnings	35,223	41,471	49,290	55,280	63,379	70,306
Less: common stock held in treasury	6,004	5,965	10,974	14,388	19,033	19,780
Total shareowners equity	*31,813*	*37,871*	*39,318*	*43,319*	*42,511*	*50,588*
Total liabilities and shareholders' equity	*53,317*	*58,025*	*70,556*	*80,954*	*84,912*	*94,682*

research and development, interest, and selling, general, and administrative expenses.

Table 2.5 presents JNJ's balance sheet for 2004–2009. Total assets are divided into current and fixed assets. Again, the current assets are those assets that can be converted into cash in a year or less; fixed assets such as land cannot be turned quickly into cash. The liabilities section is separated into liabilities and

Table 2.6 Income statement for Johnson & Johnson corporation ($ million)

(Dollars in millions except per share figures)	2004	2005	2006	2007	2008	2009
Sales to customers	$47,348	$50,514	$53,324	$61,095	$63,747	$61,897
Cost of products sold	13,422	13,954	15,057	17,751	18,511	18,447
Gross profit	33,926	36,560	38,267	43,344	45,236	43,450
Selling, marketing, and administrative expenses	15,860	16,877	17,433	20,451	21,490	19,801
Research expense	5,203	6,312	7,125	7,680	7,577	6,986
Purchased in-process research and development	18	362	559	807	181	–
Interest income	−195	−487	−829	−452	−361	–90
Interest expense, net of portion capitalized	187	54	63	296	435	451
Other (income) expense, net	15	−214	−671	1,279	−1,015	547
	21,088	22,904	23,680	30,061	26,307	27,695
Earnings before provision for taxes on income	12,838	13,656	14,587	13,283	16,929	15,755
Provision for taxes on income	4,329	3,245	3,534	2,707	3,980	3,489
Net earnings	8,509	10,411	11,053	10,576	12,949	12,266
Basic net earnings per share	$2.87	$3.50	$3.76	$3.67	$4.62	$4.45
Diluted net earnings per share	$2.84	$3.46	$3.73	$3.63	$4.57	$4.40

stockholders' equity. GM's liabilities include long-term debt, current liabilities such as accounts payable, and deferred taxes. Stockholders' equity consists of common stock and paid-in surplus, preferred stock, retained earnings, and other adjustments. Note that *total assets equal the sum of total liabilities and equity*.

Table 2.6 displays JNJ's income statement for both 2004–2009, showing the firm's profit after expenses are subtracted from revenues. To determine gross profits, the costs of goods sold are subtracted from net sales. Operating income is then calculated by deducting selling and administrative expenses, debt amortization, and depreciation. Earnings before interest and taxes (EBIT) are next obtained by adding other income to operating income, and interest is subtracted to get earnings before taxes (EBT). Finally, net income is obtained by subtracting the provision for income taxes and adding earnings in unconsolidated subsidiaries and associates. Both earnings per share and dividends per share also are reported.

Financial Ratio Analysis

To help them analyze balance sheets and income statements, financial managers construct various financial ratios. There are five basic types of these ratios: leverage ratios, activity ratios, liquidity ratios, profitability ratios, and market value ratios.

Table 2.7 Financial and market ratio calculations for JNJ, 2009 (dollars amounts in millions)

1.	$\text{Current ratio} = \frac{\text{Current Assets}}{\text{Current Liabilities}} = \frac{\$39{,}451}{\$21{,}731} = 1.82$ (liquidity ratio)
2.	$\text{Inventory turnover} = \frac{\text{Cost of Goods sold}}{\text{inventories}} = \frac{\$18{,}447}{\$5{,}180} = 3.561$ (activity ratio)[a]
3.	$\text{Total debt to total asset ratio} = \frac{\text{Total debt}}{\text{Total assets}} = \frac{\$42{,}670}{\$94{,}682} = 0.451$ (leverage ratio)
4.	$\text{Net profit margin} = \frac{\text{Net income}}{\text{Net sales}} = \frac{\$12{,}266}{\$61{,}897} = 0.198$ (profitability ratio)
5.	$\text{Return of total asset} = \frac{\text{Net income}}{\text{Total assets}} = \frac{\$12{,}266}{\$94{,}682} = 0.129$ (profitability ratio)
6.	$\text{Price / earnings ratio} = \frac{\text{Price per share}}{\text{earnings per share}} = \frac{\$64.41}{\$4.45} = 14.474$ (market value ratio)
7.	$\text{Payout ratio} = \frac{\text{dividend per share}}{\text{earnings per share}} = \frac{\$1.93}{\$4.45} = 0.434$ (market value ratio)

[a]From 10 K
From Compustat

Let us use Merck (MRK) and Johnson & Johnson data to calculate a number of these ratios and discuss their significance.

Table 2.7 shows how financial ratios are derived from the 2009 balance sheet and income statement. The information for the current ratio comes from the assets side of the balance sheet. The ratio for JNJ is 1.82, which means that 1 dollar in current liabilities is matched by 1.82 dollars in current assets. To calculate the inventory turnover ratio, we use cost of goods sold from the income statement and inventories from the current assets in the balance sheet. The resulting ratio (3.561) reveals how often the average value of goods in inventory was sold in 2009.

The total debt to total assets ratio is derived from the balance sheet; it indicates that about 45.1 % of JNJ's assets are financed by debt. Data to calculate the net profit margin come from the income statement. JNJ's profit margin is .198, that is, about 20 cents out of every dollar of sales is profit (net income). ROI (return on investment) is more accurately described as return on total assets; it is calculated from information in both the income statement and the balance sheet. The resulting figure for JNJ is 12.9 %.

The price/earnings (P/E) ratio is calculated by taking the price per share divided by the earnings per share (EPS). Although the price per share does not appear in the balance sheet or the income statement, it can be found in stock reports in newspapers. The EPS is then found by dividing net income by the number of common shares. (The ratio cannot be calculated if the firm experienced losses.) The P/E ratio for JNJ is 14.474.

Finally, the payout ratio is calculated by dividing the price of the stock by the dividends per share (DPS). DPS is the value of dividends paid out divided by the number of shares of common stock. This ratio reveals that JNJ paid out about 0.434 % of its earnings in dividends.

The seven ratios discussed in Table 2.7 for both Johnson & Johnson and Merck during 2004–2009 are presented in Table 2.8 following the method discussed in Appendix 2. Line charts in terms of data presented in Table 2.8 are presented in Figs. 2.17, 2.18, 2.19, 2.20, 2.21, 2.22 and 2.23. By using these seven graphs, we now compare the financial ratios of JNJ to those of Merck.

Table 2.8 Seven key financial ratios for JNJ and Merck

	Current ratio		Total debt to total assets ratio		Inventory turnover ratio		Return on total assets		Net profit margin		Payout ratio		Price earnings ratio	
Year	JNJ	MRK	JNJ	MRK	JNJ	MRK	JNJ	MRK	JNJ	MRK	JNJ	MRK	JNJ	MRK
1990	1.778	1.332	.485	.422	2.438	1.850	0.120	0.222	.102	.232	.381	.442	20.918	19.709
1991	1.835	1.532	.465	.389	2.321	1.796	0.138	0.223	.117	.247	.351	.434	26.082	30.328
1992	1.582	1.216	.557	.484	2.427	1.661	0.086	0.179	.075	.205	.570	.558	32.372	25.218
1993	1.624	.973	.535	.421	2.450	1.522	0.145	0.109	.126	.206	.369	.572	16.378	18.382
1994	1.566	1.270	.537	.378	2.359	3.323	0.128	0.137	.127	.200	.362	.488	17.548	16.019
1995	1.809	1.515	.485	.353	2.424	3.958	0.134	0.140	.128	.200	.344	.473	22.984	24.306
1996	1.807	1.600	.450	.355	2.517	4.375	0.144	0.160	.134	.196	.337	.462	22.926	24.883
1997	1.999	1.475	.416	.407	2.427	5.211	0.153	0.179	.146	.195	.344	.454	26.670	27.676
1998	1.364	1.685	.459	.426	2.306	5.546	0.116	0.165	.129	.195	.427	.448	36.949	33.447
1999	1.771	1.285	.434	.494	2.353	6.128	0.142	0.165	.152	.180	.355	.446	31.083	26.768
2000	2.164	1.375	.391	.471	2.475	7.340	0.153	0.171	.165	.169	.359	.426	30.453	31.630
2001	2.296	1.123	.358	.491	2.719	8.446	0.147	0.165	.172	.153	.361	.433	31.604	18.491
2002	1.683	1.199	.424	.474	2.791	9.113	0.162	0.150	.182	.138	.361	.448	24.414	17.858
2003	1.710	1.205	.427	.492	2.991	1.053	0.149	0.168	.172	.304	.382	.478	21.347	15.148
2004	1.962	1.147	.396	.510	3.082	1.534	0.159	0.137	.180	.248	.382	.573	22.098	12.267
2005	2.485	1.582	.344	.518	3.080	1.873	0.179	0.103	.206	.210	.364	.721	17.171	15.076
2006	1.199	1.197	.424	.533	2.911	2.194	0.156	0.099	.208	.196	.386	.749	17.559	21.373
2007	1.510	1.227	.446	.566	2.995	2.263	0.130	0.068	.173	.135	.442	1.011	18.174	38.483
2008	1.649	1.348	.482	.545	3.086	1.881	0.152	0.165	.203	.327	.388	.416	12.950	8.306
2009	1.820	1.805	.451	.373	3.041	.714	0.129	0.115	.198	.470	.434	.279	14.474	6.444

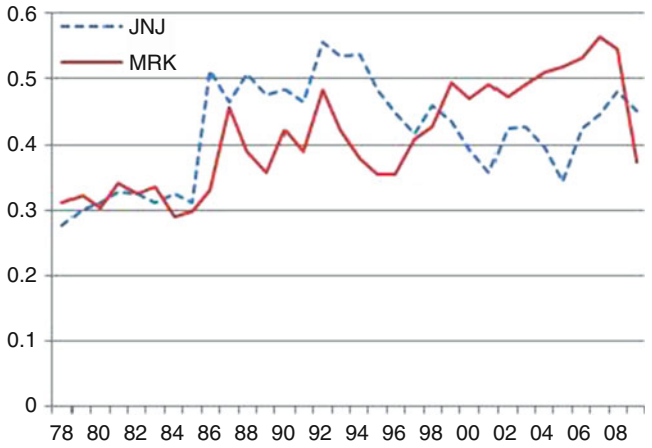


Fig. 2.17 Current ratio for JNJ and MRK

As mentioned above, there are five basic types of financial ratios. *Liquidity ratios* measure how quickly or effectively the firm can obtain cash. If the bulk of a firm's assets are fixed (such as land), the firm may not be able to obtain enough cash to finance its operations. The *current ratio*, defined as current assets divided by current liabilities, is used to gage the firm's ability to meet current obligations. If the firm's current assets do not significantly exceed current liabilities, the firm may not be able to pay current bills, because although current assets are expected to generate cash within 1 year, current liabilities are expected to use cash within that same 1-year period. In Fig. 2.17, this ratio is graphed for both Johnson & Johnson and Merck for the years 1990–2009. As the figure reveals, JNJ had a higher current ratio almost the entire time with the exception of 1998 when its current ratio was1.364, while Merck's was 1.685.

Leverage ratios measure how much of the firm's operation is financed by debt. Although some debt is expected, too much debt can be a sign of trouble. One indicator of how much debt the firm has incurred is the ratio of total debt to total assets, which measures the percentage of total assets financed by debt. Figure 2.18 shows that Johnson & Johnson had a greater share of its assets financed by debt than did Merck over most of the 1977–2009 periods. This fact is not necessarily a reason for concern unless Johnson & Johnson's leverage ratio was too high in absolute terms. The general trend shows that both firms increased their debt during the period, particularly from 1985 to 1998, and that a sharp increase occurred from 1987 to 1992.

Activity ratios measure how efficiently the firm is using its assets. Figure 2.19 graphs the *inventory turnover* ratio for each firm; it is found by dividing cost of goods sold by average inventory. This ratio measures how quickly a firm is turning over its inventories. A high ratio usually implies efficiency because the firm is selling inventories quickly. This ratio varies greatly with the line of business, however. A supermarket must have a high turnover ratio because it is dealing

Fig. 2.18 Total debt to total assets ratio for JNJ and MRK

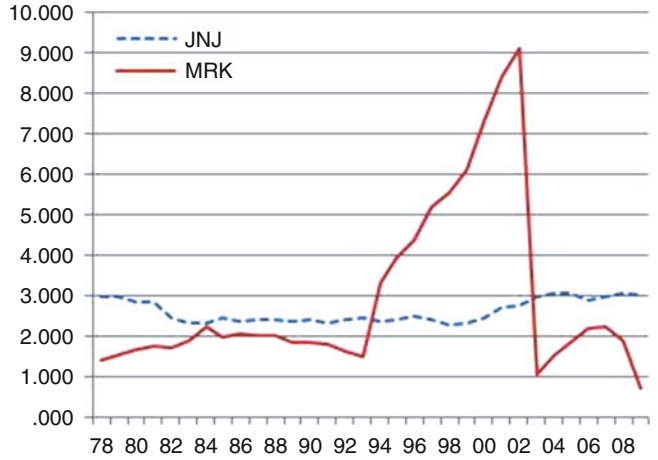


Fig. 2.19 Inventory turnover for JNJ and MRK

with perishable goods; in contrast, a jewelry store selling diamonds has a much lower turnover ratio. The seasonality of the product must also be considered. Auto dealers have high inventories in the fall, when the new autos arrive, and lower inventories in other seasons. On the other hand, Christmas tree dealers have rather low inventories in August!

Profitability ratios measure the profitability of the firm's operations. One of these ratios is the *return on total assets,* defined as net income divided by total assets. This ratio, often abbreviated ROA, measures how much the company has earned on its total investment of financial resources. Looked at in another way, it measures how well the firm used funds, regardless of how the firm's assets are

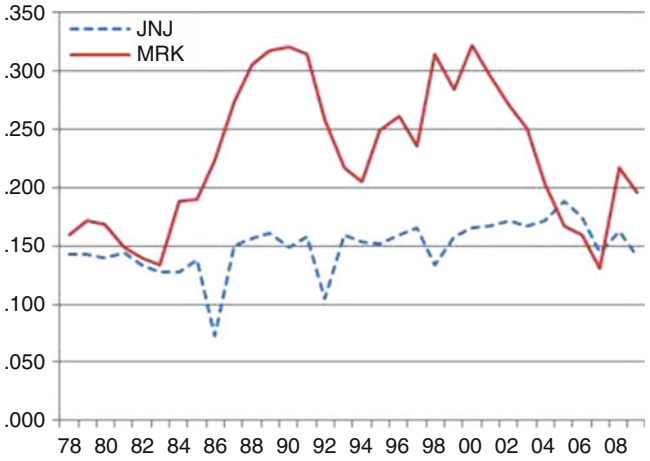


Fig. 2.20 Return on total assets for JNJ and MRK

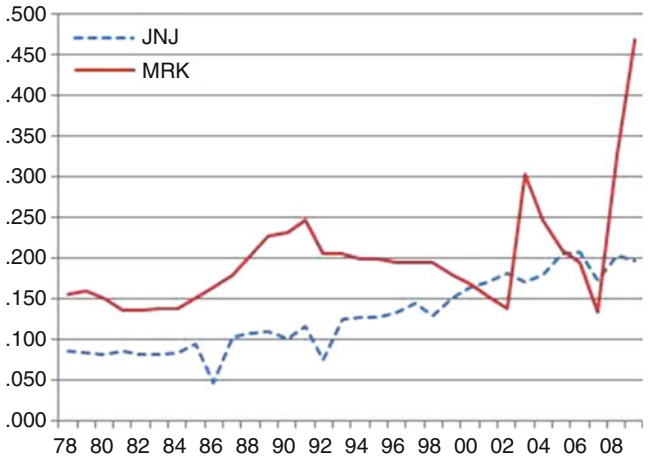


Fig. 2.21 Net profit margin for JNJ and MRK

divided into fixed and current assets. As Fig. 2.20 suggests, Merck had a higher ROA than Johnson & Johnson until 2004–2007.

The *net profit margin,* defined as net income divided by net sales, is another measure of profitability. This ratio gages the percentage of sales revenue that consists of profit. This ratio varies for different industries; a successful supermarket might have a ratio of 20 %, whereas most manufacturing firms tend to have ratios around 8 %. Although many Americans believe that corporations make a profit of 25 cents or more on each dollar of sales, the average net profit ratio for the *Fortune* 500 industrial firms in 1981 was 4.6 %. The net profit margins for Merck and Johnson & Johnson are presented in Fig. 2.21.

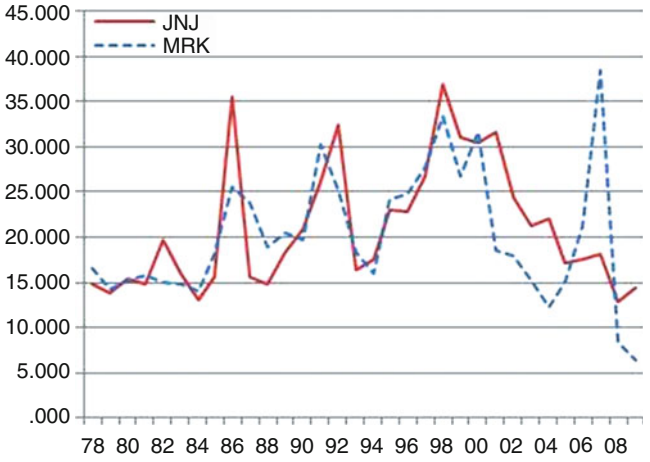


Fig. 2.22 Payout ratio for JNJ and MRK

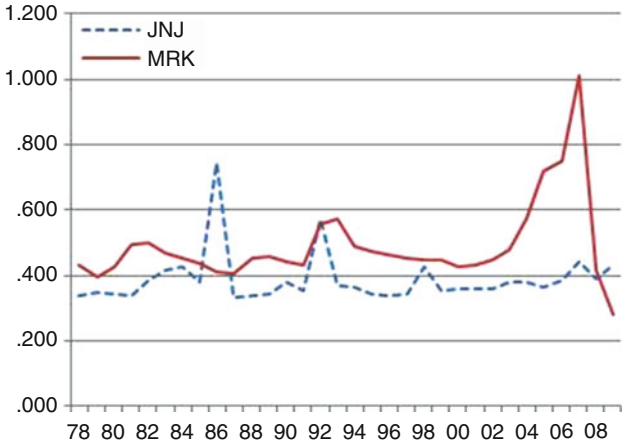


Fig. 2.23 Price/earnings ratio for JNJ and MRK

An indirect profitability indicator is the *payout ratio*, which measures the proportion of current earnings paid out in dividends. This ratio, which is expressed as dividends per share divided by earnings per share, can fluctuate widely because of the variability in earnings per share. . The reason, for example, why the payout ratio was so high for Merck in 1981 is that earnings per share were low. The payout ratios for Merck and Johnson & Johnson are illustrated in Fig. 2.22.

Market value ratios measure how stock price per share is related to either earnings per share or book value per share. The *price/earnings ratio,* or P/E, is shown in Fig. 2.23. This ratio, defined as the price per share of a stock divided by the earnings per share, is usually reported in stock quotations in newspapers such as

the *Wall Street Journal* every day. However, you should be careful in looking at P/E ratios because a high ratio can be the result of low earnings. This seems to have been the case for Merck in 2007. Moreover, firms calculate earnings per share differently, making comparisons of P/E ratios between firms difficult or even misleading.

Chapter 3
Frequency Distributions and Data Analyses

Chapter Outline

Key Terms

Grouped data	Histograms
Raw (nongrouped) data	Stem-and-leaf displays
Frequency	Frequency polygon
Frequency table	Cumulative frequency polygon
Frequency distribution	Pie chart
Cumulative frequencies	Lorenz curve
Relative frequency	Gini coefficient
Cumulative relative frequency	Absolute inequality

3.1 Introduction

Using the tabular and graphical methods discussed in Chap. 2, we will now develop two general ways to describe data more fully. We discuss first the tally table approach to depicting data frequency distributions and then three other kinds of frequency tables. Next, we explore alternative graphical methods for describing frequency distributions. Finally, we study further applications for frequency distributions in business and economics.

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_3,

Table 3.1 Student exam scores

Score	Tallies	Frequency
44	/	1
56	/	1
59	/	1
65	/	1
68	/	1
69	/	1
71	/	1
73	/	1
75	/	I
77	/	1
78	//	2
80	/	1
84	///	3
90	//	2
91	/	1
97	/	1
Total		20

3.2 Tally Table for Constructing a Frequency Table

Before conducting any statistical analysis, we must organize our data sets. One way to organize data is by using a tally table as a worksheet for setting up a frequency table. To set up a tally table for a set of data, we split the data into equal-sized classes in such a way that each observation fits into one and only one class of numbers (i.e., the classes are mutually exclusive). Sometimes data are reported in a frequency table with class intervals given but with actual values of observations in the classes unknown; data presented in this manner are called *grouped data*. The analyst assigns each data point to a class and enters a tally mark made by that class. Let's see how this works.

Example 3.1 Tallying Scores from a Statistics Exam. Suppose a statistics professor wants to summarize how 20 students performed on an exam. Their scores are as follows: 78, 56, 91, 59, 78, 84, 65, 97, 84, 71, 84, 44, 69, 90, 73, 77, 80, 90, 68, and 75. Data in this form are called *nongrouped data* or *raw data*. We can use a tally table like Table 3.1 to list the number of occurrences, of *frequency*, of each score. A corresponding diagram is shown in Fig. 3.1.

This table presents nongrouped data, and no pattern emerges from them. As an alternative, the data can be grouped into classes by letter grade. If the professor uses a straight grading scale, the classes might be 90–99, 80–89, 70–79, 60–69, and below 60. After establishing the classes, the professor counts scores in each class and records these numbers to obtain a tally sheet, as shown in Table 3.2 and Fig. 3.2.

Note that each observation is included in one and only one class. The tallies are counted, and a *frequency table* is constructed as shown in Table 3.3, where letter grades are assigned to each class.

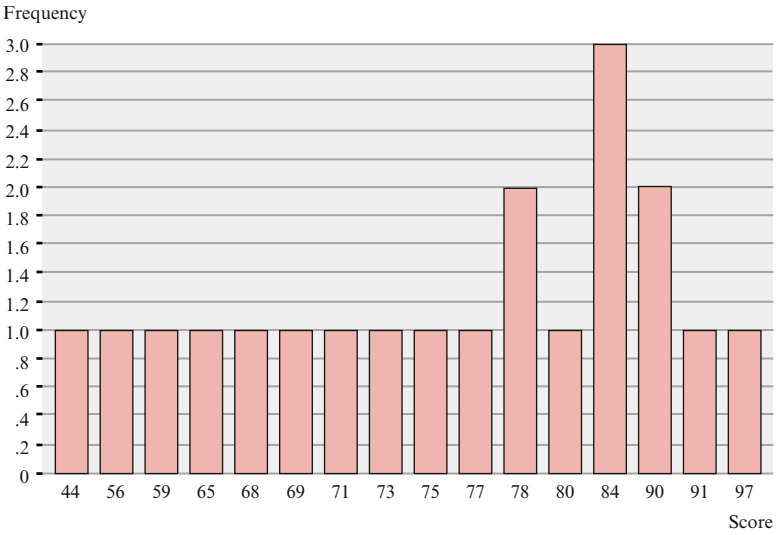


Fig. 3.1 Bar graph for nongrouped student exam scores given in Table 3.1

Table 3.2 Tally table for statistics exam scores

Class	Tally	Frequency
Below 60	///	3
60–69	///	3
70–79	//////	6
80–89	////	4
90–99	////	4
		20

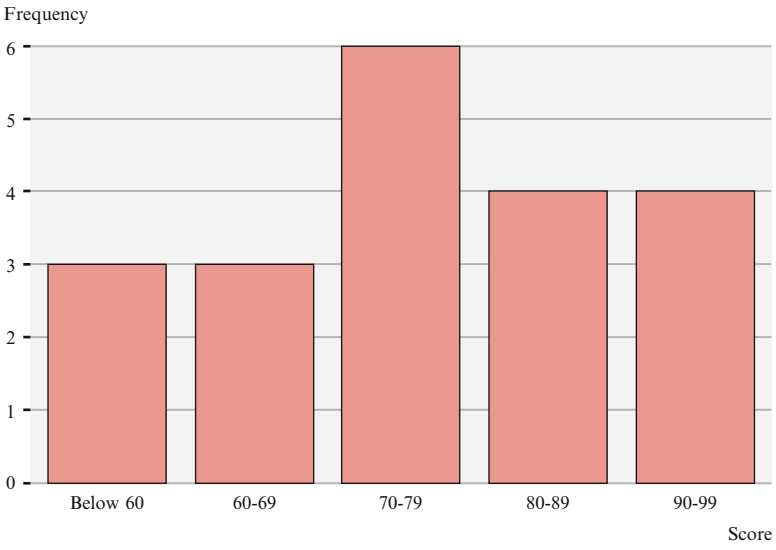


Fig. 3.2 Bar graph for grouped student exam scores given in Table 3.2

Table 3.3 Frequency table for statistics exam scores

Class	Grade	Frequency
Below 60	F	3
60–69	D	3
70–79	C	6
80–89	B	4
90–99	A	4
		20

Table 3.4 Student GPAs: raw data

1.2	3.9	1.9
3.8	2.4	2.7
2.3	2.3	2.6
0.7	3.1	3.7
3.6	2.9	4.0
2.2	2.7	1.2
1.9	0.8	1.8
2.1	0.3	2.4
3.1	3.2	3.2
0.8	3.1	3.6

Table 3.5 Student GPAs: tally table and frequency distribution

Range	Tallies	Frequency
Below 1.5	//////	6
1.5–1.9	///	3
2.0–2.4	//////	6
2.5–2.9	////	4
3.0–3.4	/////	5
3.5–4.0	//////	6
Total		30

Example 3.2 A Frequency Distribution of Grade Point Averages. Suppose that there are 30 students in a classroom and that they have the grade point averages listed in Table 3.4. A tally table is constructed, in which classes are (arbitrarily) defined at every half-point and each tally marked next to a particular class accounts for one data entry. The entries are then counted to obtain a *frequency distribution*, as shown in Table 3.5. A frequency distribution simply shows how many observations fall into each class. We will discuss this concept in further detail in the next section.

Generally, a data set should be divided into 5–15 classes. Having too few or too many classes gives too little information. Imagine a frequency distribution with only two classes: 0.0–2.0 and 2.1–4.0. With such broadly defined classes, it is difficult to distinguish among GPAs. Similarly, if the class interval were only one-tenth of a point, the large number of classes, each with only one or a few tallies, would make summarizing the data almost impossible.

Table 3.6 T-bill interest rates, 1990–2009

Class (%)	Tallies	Frequency
0–1.49	////	4
1.50–3.49	/////	5
3.50–5.49	/////////	9
5.50–6.49	/	1
6.50 and greater	/	1
Total		20

In the GPA example, it was relatively easy to construct the classes because GPA cutoffs were used. However, in most examples, there are no natural dividing lines between classes. The following guidelines can be used to construct classes:

1. Construct from 5 to 15 classes. This step is the most difficult, because using too many classes defeats the purpose of grouping the data into classes, whereas having too few classes limits the amount of information obtained from the data. As a general rule, when the range and number of observations are large, more classes can be defined. Fewer classes should be constructed when the number of observations is only around 20 or 30.
2. Make sure each observation falls into only one class. This can often be accomplished by defining class boundaries in terms of several decimal places. If the percentage return on stocks is carried to one decimal place, for example, then defining the classes by using two decimal places will ensure that each observation falls into only one class.
3. Try to construct classes with equal class intervals. This may not be possible, however, if there are outlying observations in the data set.

Example 3.3 A Frequency Distribution of 3-Month Treasury Bill Rates. Table 3.6 presents another example, and here the data presented are the interest rates on 3-month treasury bills (T-bills) from 1990 to 2009. (T-bills are debt instruments sold by the US government to finance its budgetary needs.) The annual data for interest rates (average daily rates for a year) are taken from Economic Report of the President, January 2009.

As we have noted, a frequency distribution gives the total number of occurrences in each class. In the next chapter, we will talk about using a frequency distribution to present data.

By setting up a tally table and a frequency table, we can scrutinize data for errors. For example, if the data value 123 appears in a column for the rate in the T-bill example, a mistake has clearly been made – one that could be due to a missing decimal point. Probably, the data point could be 12.3 % instead, which makes more sense because it is in the range of the other data points. Data should also be checked for accuracy. Otherwise, invalid conclusions could be reached.

Table 3.7 Cumulative frequency table for grade distribution

Class	Grade	Frequency	Cumulative frequency
Below 60	F	3	3
60–69	D	3	6
70–79	C	6	12
80–89	B	4	16
90–99	A	4	20

3.3 Three Other Frequency Tables

In this section, using the frequency table discussed in the Sect. 3.2, we move ahead to cumulative frequency tables, relative frequency tables, and relative cumulative frequency tables.

Example 3.4 Frequency Distributions for Statistics Exam Scores. Suppose that for the data listed in Table 3.3, the professor wants to know how many students receive a C or below, the proportion of students who receive a B, and the proportion of students who receive a D or an F. To obtain this information, she calculates cumulative, relative, and cumulative relative frequencies.

By constructing *cumulative frequencies*, the professor determines the number of students who scored in a particular class *or* in one of the classes before it (Table 3.7). Obviously, the cumulative frequency for the first class is the frequency itself (3): there are no classes before it. The cumulative frequency for the second class is calculated by taking the frequency in the first class and adding it to the frequency in the second class (3) to arrive at a cumulative frequency of 6. This means that 6 students were in the first two classes. Then 6 is added to the frequency of the third class (6) to derive a cumulative frequency of 12. Thus, 12 students scored a C or a worse grade. The remaining cumulative frequencies are calculated in a similar manner. Note that the cumulative observation in the last class equals the total number of sample observations, because all frequencies have occurred in that class or in previous classes.

Another important concept is the *relative frequency*, which measures the proportion of observations in a particular class. It is calculated by dividing the frequency in that class by the total number of observations. For the data summarized in Table 3.7, the relative frequency for both the first and second classes is 0.15, and the relative frequencies for the remaining three classes are 0.30, 0.20, and 0.20, respectively, as shown in Table 3.8. The sum of the relative frequencies always equals 1.

This table indicates that 15 % of the class received an F, 15 % a D, 30 % a C, and so on. The professor can calculate the cumulative relative frequency for any class by adding the appropriate relative frequencies. *Cumulative relative frequency* measures the percentage of observations in a particular class and all previous classes. Thus, if she wants to determine what percentage of the students scored below a B, our conscientious professor can add the relative frequencies associated with grades C, D, and F to arrive at 60 %.

Table 3.8 Relative frequency table for grade distribution

Class	Grade	Relative frequency	Cumulative relative frequency
Below 60	F	0.15	0.15
60–69	D	0.15	0.30
70–79	C	0.30	0.60
80–89	B	0.20	0.80
90–99	A	0.20	1.00

Table 3.9 Current ratio for JNJ and MRK

Year	JNJ	MRK
1990	1.778	1.332
1991	1.835	1.532
1992	1.582	1.216
1993	1.624	0.973
1994	1.566	1.270
1995	1.809	1.515
1996	1.807	1.600
1997	1.999	1.475
1998	1.364	1.685
1999	1.771	1.285
2000	2.164	1.375
2001	2.296	1.123
2002	1.683	1.199
2003	1.710	1.205
2004	1.962	1.147
2005	2.485	1.582
2006	1.199	1.197
2007	1.510	1.227
2008	1.649	1.348
2009	1.820	1.805

Example 3.5 Frequency Distributions of Current Ratios for JNJ and MRK. The current ratios for JNJ and MRK from 1990 to 2009 are shown in Table 3.9. A frequency distribution for the current ratios of Johnson and Johnson and Merck is shown in Table 3.10. This ratio is a measure of liquidity, which (as we noted in Chap. 2) indicates how quickly a firm can obtain cash for operations. The first column defines the classes. Note that the use here of class boundaries ensures that each observation will fall into only one class.

The next column shows the frequency – that is, the number of times that an observation appears in each class. In Table 3.10, we see that JNJ experienced one current ratio between 1.0 and 1.2, seven between 1.201 and 1.700, and so on. The third column presents the cumulative frequency. Because there are 20 observations in the population, the cumulative frequency for the last class is 20.

The fourth column presents the relative frequency, which measures the percentage of observations in each class. Relative frequencies can be thought of as probabilities. For example, the probability that an observation is in the first class is 0.1.

Table 3.10 Frequency distributions of current ratios for JNJ and MRK

Class	Frequency	Cumulative frequency	Relative frequency	Cumulative relative frequency
JNJ				
1.00–1.2	1	1	0.05	0.05
1.21–1.4	1	2	0.05	0.1
1.41–1.60	3	5	0.15	0.25
1.61–1.80	6	11	0.3	0.55
1.81–2.00	6	17	0.3	0.85
2.01–2.5	3	20	0.15	1.00
Total	20		1.00	
MRK				
0.81–1.00	1	1	0.05	0.05
1.01–1.2	4	5	0.2	0.25
1.21–1.4	8	13	0.6	0.65
1.41–1.60	5	18	0.25	0.9
1.61–1.80	1	19	0.05	0.95
1.81–2.00	1	20	0.05	1.00
Total	20		1.00	

The last column indicates the cumulative relative frequency, which measures the percentage of observations in a particular class and all previous classes. The cumulative relative frequency for Merck's fourth class is calculated by adding the relative frequencies of the first four classes to arrive at 0.95. That is, 95 % of the observations occur in the first four classes. The cumulative relative frequency of the last class always equals 1, because the last class includes all the observations.

3.4 Graphical Presentation of Frequency Distribution

We have spoken before of the special effectiveness of using graphs to present data. In this section, we discuss four different graphical approaches to presenting frequency distributions.

3.4.1 Histograms

Frequency distributions can be represented on a variety of graphs. *The histogram*, which is one of the most commonly used types, is similar to the bar charts discussed in Chap. 2 except that

1. Neighboring bars touch each other.
2. The area inside any bar (its height times its width) is proportional to the number of observations in the corresponding class.

To illustrate these two points, suppose the age distribution of personnel at a small business is as shown in Table 3.11.

Table 3.11 Age distribution of personnel

Class	Frequency
20–29	3
30–39	6
40–49	7
50–59	4
60–69	1
70–79	1

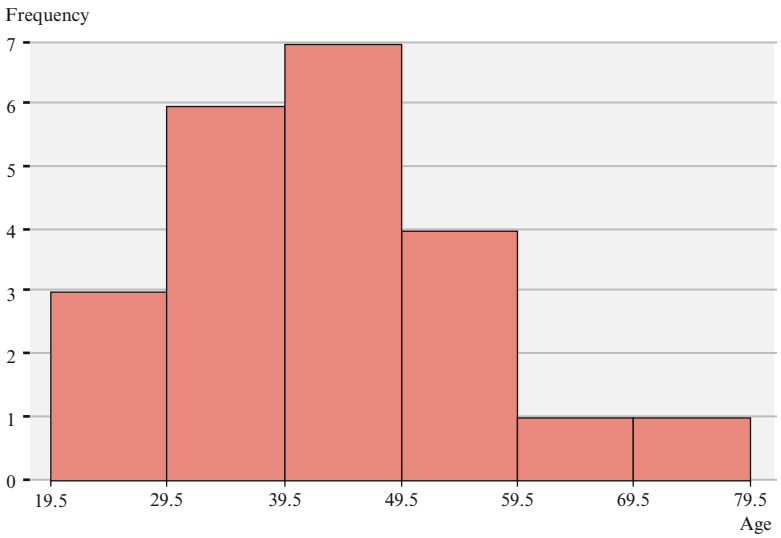


Fig. 3.3 Histogram of age distribution given in Table 3.11

To construct a histogram, we need to enter a scale on the horizontal axis. Because the data are discrete, there is a gap between the class intervals, say between 20 and 29 and 30–39. In such a case, we will use the midpoint between the end of one class and the beginning of the next as our dividing point. Between the 20–29 interval and 30–39 interval, the dividing point will be $(29 + 30)/2 = 29.5$. We find the dividing point between the remaining classes similarly.

To satisfy the second condition, we note that all five classes have an interval width of 10 years. Figure 3.3 is the histogram that reflects these data.

Drawn from the data of Table 3.10, Fig. 3.4a, b are histograms of JNJ's and MRK's current ratios for the years 1990–2009.The x-axis indicates the classes and the y-axis the frequencies. As the histograms show, MRK's current ratios have tended to fall in the 1.0–1.4 range, whereas those of JNJ show no exact pattern, but many can be found in the 1.61–2.00 range. (In Chap. 4, we will cover measures of skewness, which give us more insight into the shape of a distribution.)

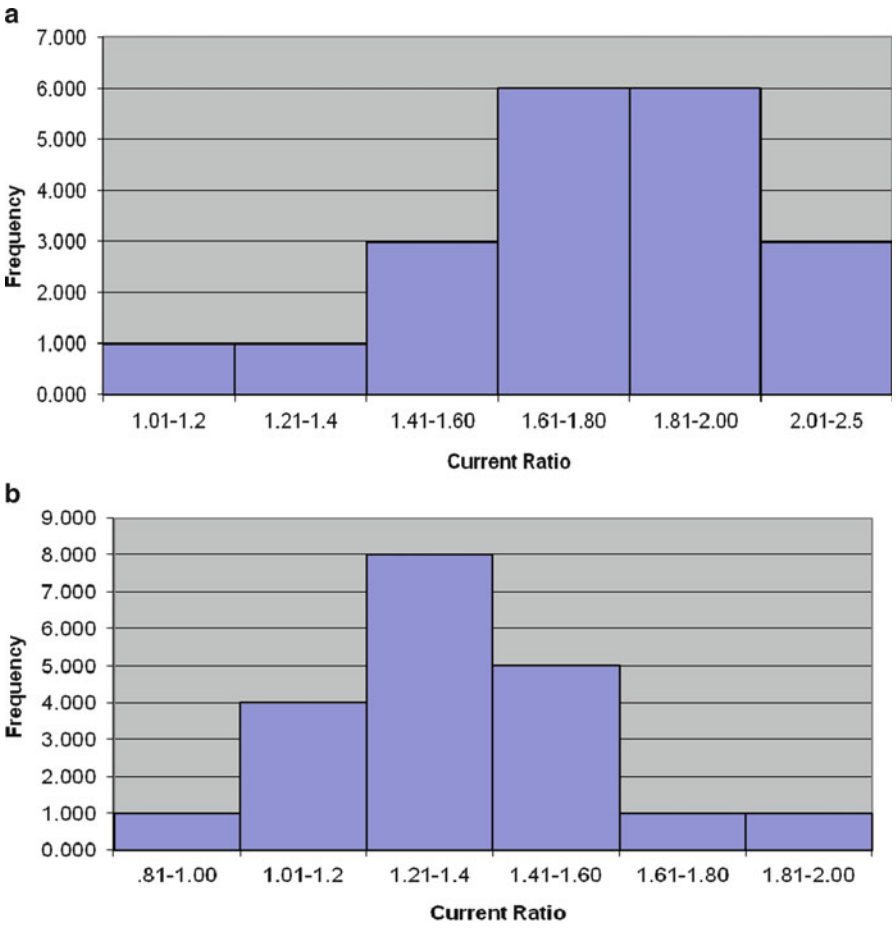


Fig. 3.4 (**a**) Frequency histogram of JNJ's current ratios as given in Table 3.10 (**b**) Frequency histogram of MRK's current ratios as given in Table 3.10

Most standard statistical software packages will construct a histogram from these data. Using MINITAB, we can specify the class width and the starting class midpoint, or we can let MINITAB select these values. The output will contain the frequency distribution as well as a graphical representation in the form of a histogram (without the bars). MINITAB will provide each class frequency next to the corresponding class midpoint (not class limits). Figure 3.5a contains the necessary MINITAB commands and the resulting output for the current ratio of MRK where the class width (CW) and the midpoint of the first class were not specified. Figure 3.5b specified CW as .2000 and the first midpoint as .905. We can use the output as it appears or use this information to construct Fig. 3.4b, which is a graphical representation of MRK's current ratios as given in Table 3.10.

a

Data Display

```
MRK
   1.332   1.532   1.216   0.973   1.27    1.515   1.6     1.475   1.685
   1.285   1.375   1.123   1.199   1.205   1.147   1.582   1.197   1.227
   1.348   1.805
```

Histogram of MRK

```
* NOTE * The character graph commands are obsolete.
```

Histogram

```
Histogram of MRK   N = 20

Midpoint        Count
     1.0            1  *
     1.1            2  **
     1.2            5  *****
     1.3            4  ****
     1.4            1  *
     1.5            3  ***
     1.6            2  **
     1.7            1  *
```

b

Histogram

```
Histogram of MRK   N = 20

Midpoint        Count
   0.905            1  *
   1.105            4  ****
   1.305            8  ********
   1.505            5  *****
   1.705            1  *
   1.905            1  *
```

Fig. 3.5 (**a**) Histogram using MINITAB, where the class width and the midpoint of the first class are not specified (**b**) Histogram using MINITAB using specified classes, where the class width is 0.2000 and the first midpoint is 0.905

Histograms can also be used to chart the companies' relative and cumulative frequencies, as shown in Figs. 3.6 and 3.7. Note the similarity between the frequency and relative frequency histograms (Figs. 3.4 and 3.6) and between the cumulative frequency and the relative cumulative frequency graphs (Figs. 3.7 and 3.8); the only difference between them is the variable on the y-axis. Note also that geometrically, the relative frequency of each class in a frequency histogram equals its area divided

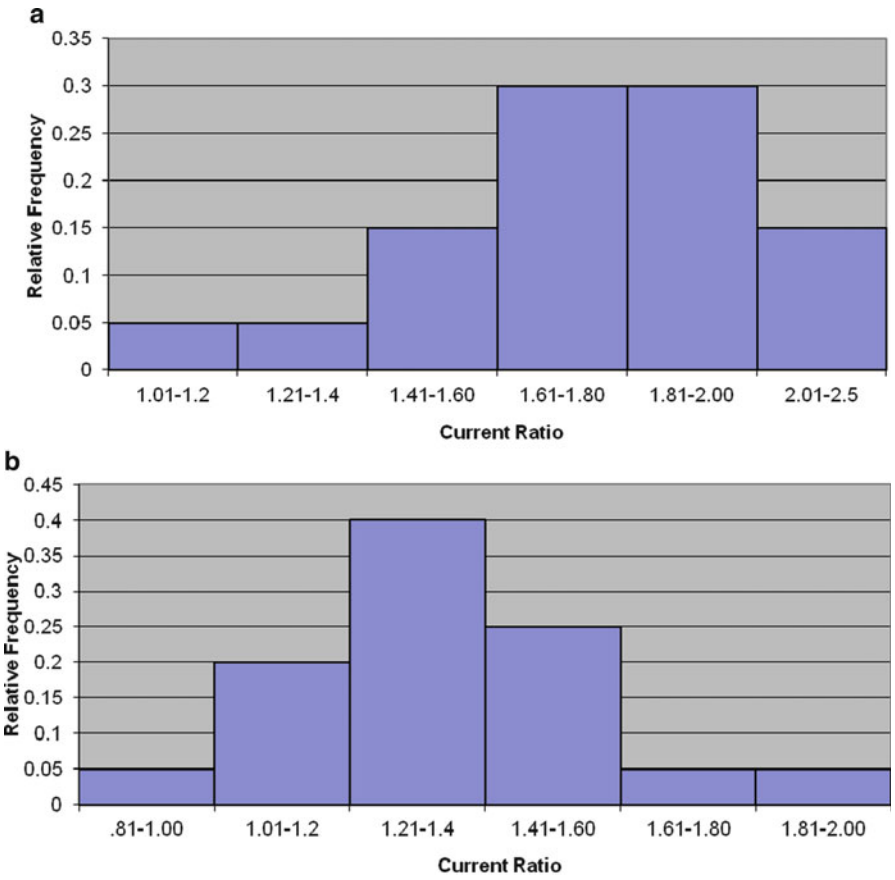


Fig. 3.6 (**a**) Relative frequency histogram of JNJ's current ratios (**b**) Relative frequency histogram of MRK's current ratios

by the total area of all the classes. For example, the area for the first class for Merck's current ratio (Fig. 3.4b) is equal to the base of the bar times its height ($0.19 \times 1 = 0.19$), and the sum of all the areas is 3.8. The relative frequency for the first class is thus $.19/3.8 = .05$.

3.4.2 Stem-and-Leaf Display

An alternative to histograms for the presentation of either nongrouped or grouped data is the stem-and-leaf display. *Stem-and-leaf displays* were originally developed by John Tukey of Princeton University. They are extremely useful in summarizing data sets of reasonable size (under 100 values as a general rule), and unlike histograms, they result in no loss of information. By this, we mean that it is possible

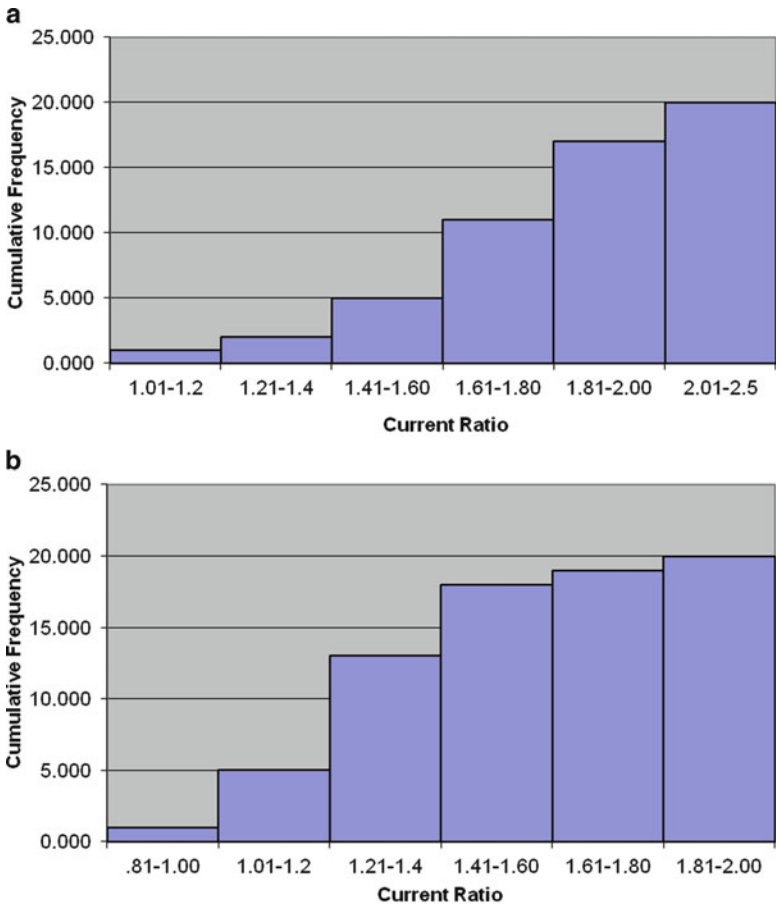


Fig. 3.7 (**a**) Cumulative Frequency histogram of JNJ's current ratios (**b**) Cumulative frequency histogram of MRK's current ratios

to reconstruct the original data set in a stem-and-leaf display, which we cannot do when using a histogram.

For example, suppose a financial analyst is interested in the amount of money spent by food product companies on advertising. He or she samples 40 of these food product firms and calculates the amount that each spent last year on advertising as a percentage of its total revenue. The results are listed in Table 3.12.

Let's use this set of data to construct a stem-and-leaf display. In Fig. 3.9, each observation is represented by a stem to the left of the vertical line and a leaf to the right of the vertical line. For example, the stems and leaves for the first three observations in Table 3.12 can be defined as

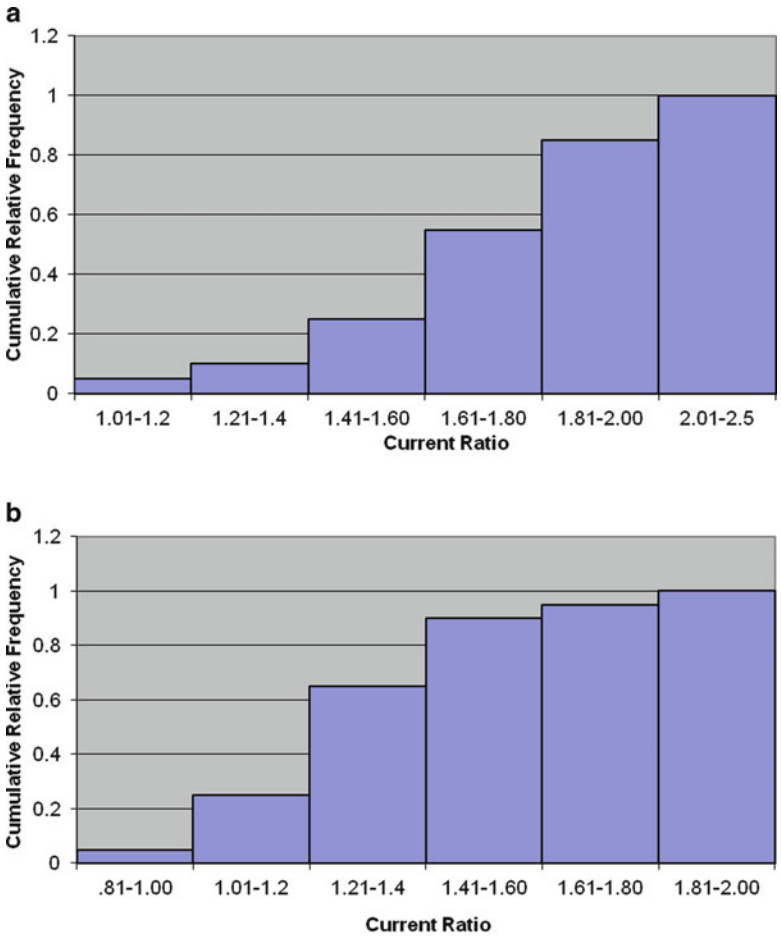


Fig. 3.8 (**a**) Cumulative relative frequency histogram of JNJ's current ratios (**b**) Cumulative relative frequency histogram of MRK's current ratios

Stem	Leaf
12	0.5
8	0.8
11	0.5

In other words, stems are the integer portions of the observations, whereas leaves represent the decimal portions.

The procedure used to construct a stem-and-leaf display is as follows:

1. Decide how the stems and leaves will be defined.
2. List the stems in a column in ascending order.

Table 3.12 Percentage of total revenue spent on advertising

Company	Percentage	Company	Percentage
1	12.5	21	6.4
2	8.8	22	7.8
3	11.5	23	8.5
4	9.1	24	9.5
5	9.4	25	11.3
6	10.1	26	8.9
7	5.3	27	6.6
8	10.3	28	7.5
9	10.2	29	8.3
10	7.4	30	13.8
11	8.2	31	12.9
12	7.8	32	11.8
13	6.5	33	10.4
14	9.8	34	7.6
15	9.2	35	8.6
16	12.8	36	9.4
17	13.9	37	7.3
18	13.7	38	9.5
19	9.6	39	8.3
20	6.8	40	7.1

Fig. 3.9 Stem-and-leaf display for advertising expenditure

Stems	Leaves	Frequency
5	3	1
6	4 5 6 8	4
7	1 3 4 5 6 8 8	7
8	2 3 3 5 6 8 9	7
9	1 2 4 4 5 5 6 8	8
10	1 2 3 4	4
11	3 5 8	3
12	5 8 9	3
13	7 8 9	3
Total		40

3. Proceed through the data set, placing the leaf for each observation in the appropriate stem row. (You may want to place the leaves of each stem in increasing order.)

The percentage of revenues spent on advertising by 40 production firms listed in Table 3.12 is represented by a stem-and-leaf diagram in Fig. 3.9. From this diagram, we observe that the minimum percentage of advertising spending is 5.3 % of total revenue, the maximum percentage of advertising spending is 13.9 %, and the largest group of firms spends between 9.1 % and 9.8 % of total revenue on advertising. Also, the 7 leaves in stem row 7 indicate that 7 firms' advertising spending is at least 7 % but less than 8 %. The 3 leaves in stem row 13 tell us at a

Data Display

```
ADV EXP
      12.5        8.8      11.5      9.1      9.4     10.1      5.3     10.3
      10.2        7.4       8.2      7.8      6.5      9.8      9.2     12.8
      13.9       13.7       9.9      6.8      6.4      7.8      8.5      9.5
      11.3        8.9       6.6      7.5      8.3     13.8     12.9     11.8
      10.4        7.6       8.6      9.4      7.3      9.5      8.3      7.1

MTB   >   STEM   AND   LEAF   USING    'ADV   EXP'
```

Character Stem-and-Leaf Display

```
Stem-and-leaf of ADV EXP N = 40
Leaf Unit = 0.10


   1        5 3
   5        6 4568
  12        7 1345688
  19        8 2335689
  (8)       9 12445589
  13       10 1234
   9       11 358
   6       12 589
   3       13 789
```

Fig. 3.10 Stem-and-leaf diagram for advertising expenditure using MINITAB

glance that 3 firms spend more than 13 % of total revenue on advertising. A MINITAB version of the stem-and-leaf diagram generated by these data is shown in Fig. 13.10. A stem-and-leaf diagram is presented in the last portion of Fig. 3.10. In the first column of this diagram, (8) represents the total observation in the middle group with a stem of 9; 1, 5, 12, and 19 represent the cumulative frequencies from the first group up to the fourth group; and 3, 6, 9, and 13 represent the cumulative frequencies from the ninth group up to the sixth group.

3.4.3 Frequency Polygon

A *frequency polygon* is obtained by linking the midpoints indicated on the x-axis of the class intervals from a frequency histogram. A *cumulative frequency polygon* is derived by connecting the midpoints indicated on the x-axis of the class intervals from a cumulative frequency histogram. Figures 3.11 and 3.12 show the frequency polygon and the cumulative frequency polygon, respectively, for JNJ's current ratio. Although a histogram does demonstrate the shape of the data, perhaps the shape can be more clearly illustrated by using a frequency polygon.

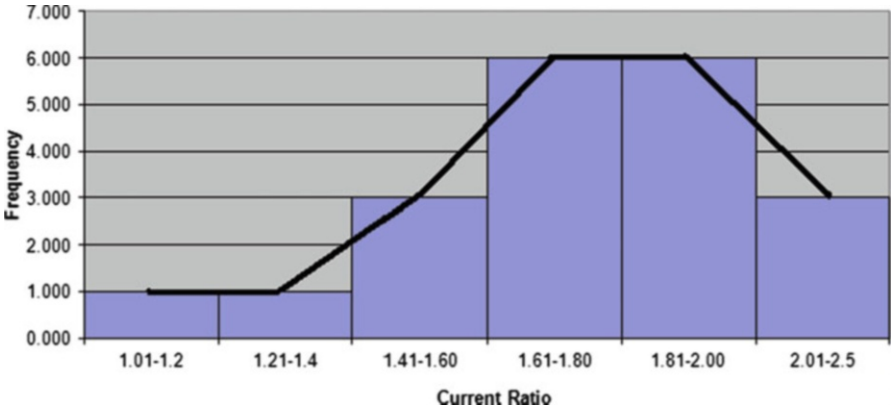


Fig. 3.11 Frequency polygon of JNJ's current ratios

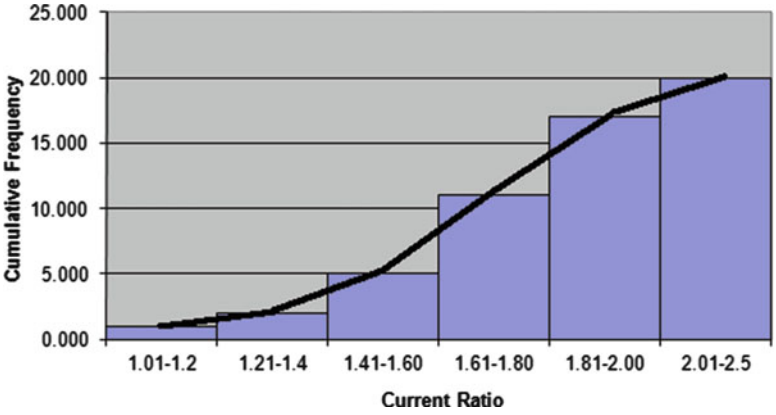


Fig. 3.12 Cumulative frequency polygon of JNJ's current ratios

3.4.4 Pie Chart

Histograms are perhaps the graphical forms most commonly used in statistics, but other pictorial forms, such as the *pie chart*, are often used to present financial and marketing data. For example, Fig. 3.13 depicts a family's sources of income. This pie chart indicates that 80 % of this family's income comes from salary.

For data already in frequency form, a pie chart is constructed by converting the relative frequencies of each class into their respective arcs of a circle. For example, a pie chart can be used to represent the student grade distribution data originally presented in Table 3.3. In Table 3.13, the arcs (in degrees) for the five slices shown in Fig. 3.14 were obtained by multiplying each relative frequency by 360°.

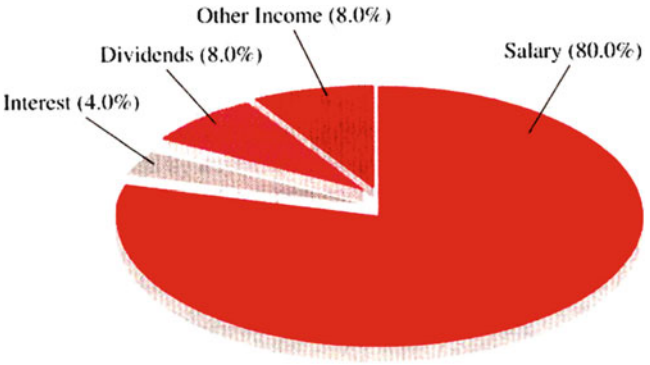


Fig. 3.13 Sources of family income

Table 3.13 Grade distribution for 20 students

Class	Frequency	Relative frequency	Arc (degrees)
Below 60	3	0.15	54
60–69	3	0.15	54
70–79	6	0.30	108
80–89	4	0.20	72
90–99	4	0.20	72
Total	20	1.00	360

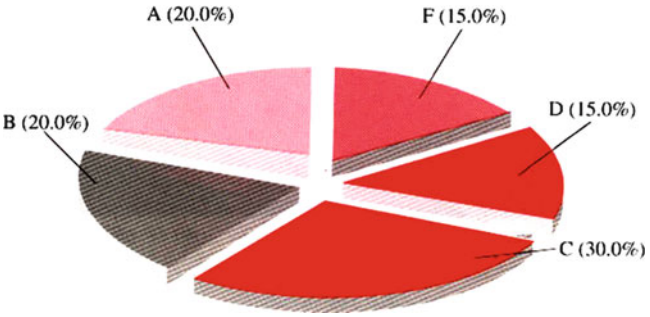


Fig. 3.14 Grade distribution pie chart

3.5 Further Economic and Business Applications

3.5.1 Lorenz Curve

The *Lorenz curve*, which represents a society's distribution of income, is a cumulative frequency curve used in economics (Fig. 3.15a). The cumulative percentage of families (ranked by income) is measured on the x-axis, and the cumulative

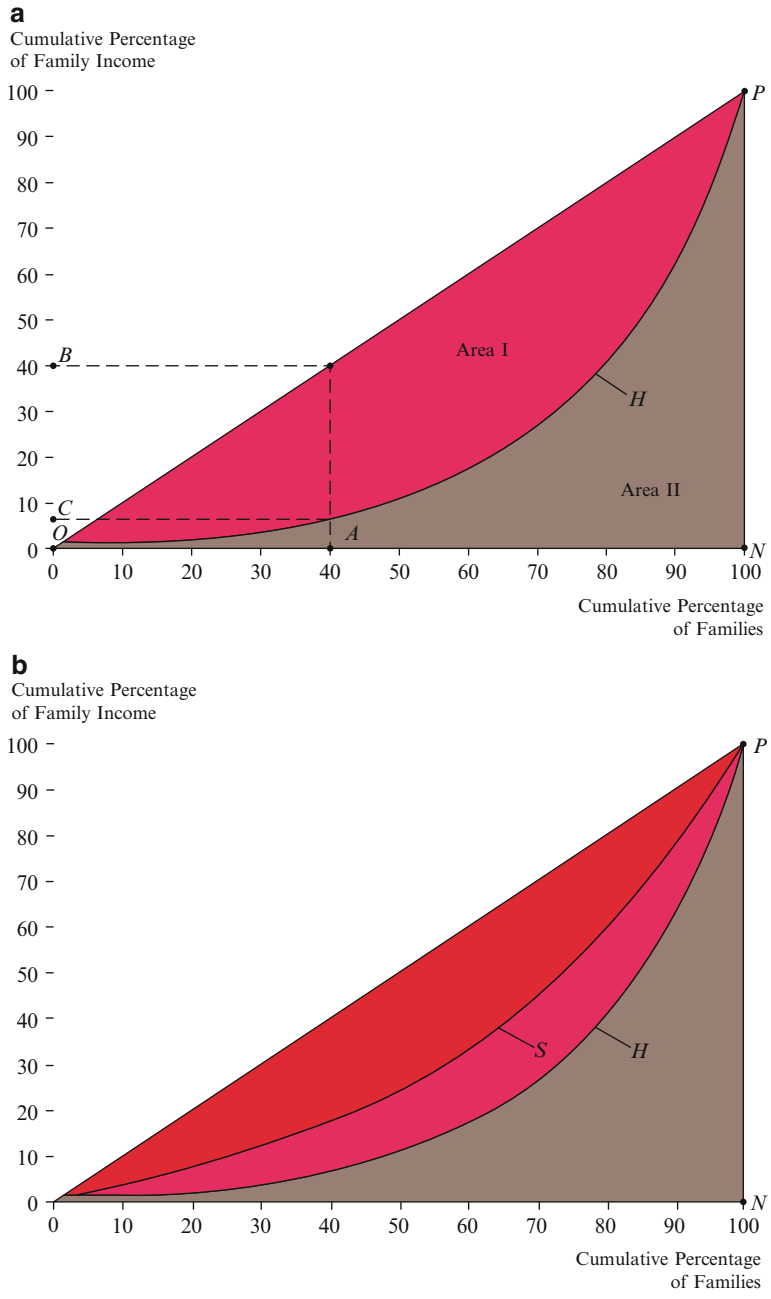


Fig. 3.15 (**a**) and (**b**) Lorenz curves

percentage of family income received is measured on the y-axis. For example, suppose there are 100 families, and each earns $100 – that is, the distribution of income is perfectly equal. The resulting Lorenz curve will be a 45° line (*OP*), because the cumulative percentage of families (e.g., 40 %) and the cumulative share of family income received are always equal.

Now suppose that one family receives 100 % of total family income – that is, the income distribution is absolutely *un*equal. The resulting Lorenz curve (*ONP*) coincides with the x-axis until point *N*, where there is a discontinuous jump to point *P*. This is because, with the exception of that single family (represented by point *N*), each family receives 0 % of total family income. Therefore, these families' cumulative share of total family income is also 0 %.

The shape the Lorenz curve is most likely to assume is curve *H*, which lies between absolute inequality and equality. This curve indicates that the lowest-income families, who comprise 40 % of families (point *A*), receive a disproportionately small share (about 7 %) of total family income (point *C*). If every family had the same income, the share going to the lowest 40 % would be represented by point *B* (40 %).

Note that with a more equitable distribution of income, the Lorenz curve is less bowed, or flatter. Curve *S* in Fig. 3.15b is the Lorenz curve after a progressive income tax is imposed. Because *S* is flatter than *H* (which is reproduced from Fig. 3.15a), we can conclude that the distribution of income (after taxes) is more nearly equal than before, as would be expected.

One way to measure the inequality of income from the Lorenz curve is to use the Gini coefficient.

$$\text{Gini coefficient for curve } H = \frac{\text{area I}}{\text{area (I + II)}}$$

The *Gini coefficient* can range from 0 (perfect equality) to 1 (*absolute inequality*, wherein one family receives all the income).

Examining Fig. 3.15b reveals that the Gini coefficient will be smaller for curve *S* than it is for curve *H*. In other words, the progressive income tax makes the distribution of income more nearly equal.

3.5.2 Stock and Market Rate of Return

Table 3.14 presents the frequency tables for the rate of return for Johnson and Johnson, Merck, and the stock market overall. (The data are drawn from Table 2.4 in Appendix 2 of Chap. 2.) Because the two firms have similar frequency distributions, we can conclude that the performances of Johnson and Johnson and Merck's stocks have been similar. However, Johnson and Johnson's highest class is

Table 3.14 Rates of return for JNJ and MRK stock and the S&P 500

Class	Frequency (years)	Cumulative frequency	Relative frequency	Cumulative relative frequency
JNJ				
−0.200 and below	4	4	0.1905	0.1905
−0.199 to 0.000	5	9	0.2381	0.4286
0.001–0.200	5	14	0.2381	0.6667
0.201–0.400	5	19	0.2381	0.9048
0.401–0.600	1	20	0.0476	0.9524
0.601–1.00	1	21	0.0476	1.0000
Total	21		1.000	
MRK				
−0.200 and below	5	5	0.2381	0.2381
−0.199 to 0.000	3	8	0.1429	0.3810
0.001–0.200	3	11	0.1429	0.5238
0.201–0.400	5	16	0.2381	0.7619
0.401–0.600	3	19	0.1429	0.9048
0.601–1.00	2	21	0.0952	1.0000
Total	21		1.000	
S&P 500 (*market*)				
−0.200 and below	1	1	0.0476	0.0476
−0.199 to 0.000	4	5	0.1905	0.2381
0.001–0.200	11	16	0.5238	0.7619
0.201–0.400	5	21	0.2381	1.0000
Total	21		1.000	

0.001–0.200, while Merck's highest classes are spread but found at −0.200 and below and at 0.201–0.400.

The stock market's overall lowest class was found at −0.200 and below, but its highest class was only 0.001–0.200. Thus, the overall market has fluctuated less than the return of the two pharmaceutical firms. And although Johnson and Johnson and Merck have a higher top class, the market suffered through fewer negative returns. Moreover, Johnson and Johnson and Merck had 9 and 8 years, respectively, of losses, while the market had only five. In other words, the pharmaceutical firms offered the potential of higher returns but also threatened the investor with a greater risk of loss.

3.5.3 Interest Rates

Histograms can be used to summarize movements in such interest rates as the prime rate and the treasury bill rate. The prime rate is the interest rate that banks charge to their best customers; treasury bills are short-term debt instruments issued by the US

Table 3.15 3-Month T-bill rate and prime rate (1990–2009)

Year	3-Month T-bill rate	Prime rate
90	7.49	10.01
91	5.38	8.46
92	3.43	6.25
93	3.00	6.00
94	4.25	7.14
95	5.49	8.83
96	5.01	8.27
97	5.06	8.44
98	4.78	8.35
99	4.64	7.99
00	5.82	9.23
01	3.39	6.92
02	1.60	4.68
03	1.01	4.12
04	1.37	4.34
05	3.15	6.19
06	4.73	7.96
07	4.35	8.05
08	1.37	5.09
09	0.15	3.25

Table 3.16 Frequency distributions of interest rates

	T-bill		Prime rate	
Class (%)	Frequency	Relative frequency	Frequency	Relative frequency
0–1.99	0	0.00	5	0.25
2–2.99	0	0.00	0	0.00
3–3.99	1	0.05	4	0.20
4–4.99	3	0.15	5	0.25
5–5.99	1	0.05	5	0.25
6–6.99	4	0.20	0	0.00
7–7.99	3	0.15	1	0.05
8–8.99	6	0.30	0	0.00
9–9.99	1	0.05	0	0.00
10–10.99	1	0.05	0	0.00
Total	20	1.00	20	1.00

government. Let us examine how these rates have moved over the period 1990–2009, as shown in Table 3.15.

As can be seen in Table 3.16 and Fig. 3.16, the prime rate is skewed to the right, with 65 % of the observations appearing in the ranges made up of the slightly higher midrange interest rates (6–6.9 %, 7–7.9 %, and 8–8.9 %). If you were to predict a future value for the prime rate, your best guess would be in the 6–9 % range. This wide range would probably not be of much use. Better methods for prediction, such as multiple regression and time series analysis, will be discussed later (Chaps. 15 and 18).

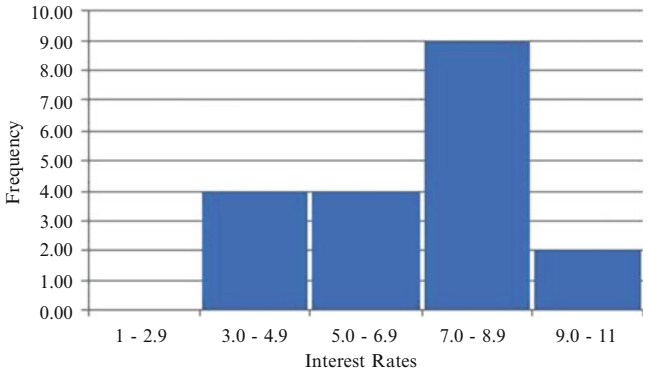


Fig. 3.16 Frequency histogram of prime lending rates given in Table 3.15

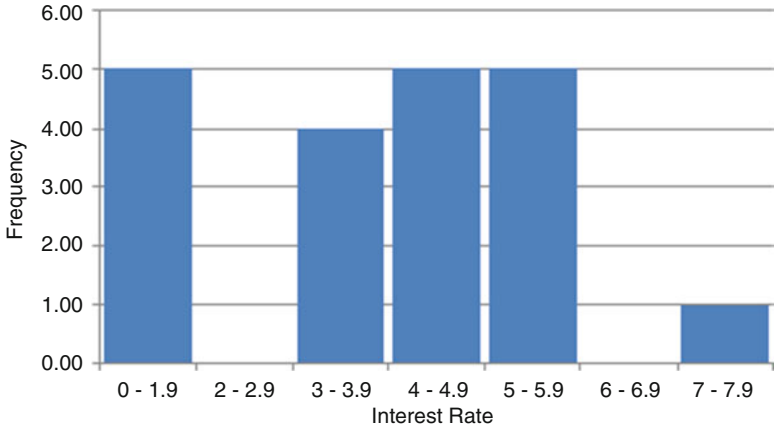


Fig. 3.17 Frequency histogram of T-bill rates given in Table 3.15

The frequency table for the treasury bill rate is shown in Table 3.16. This distribution, like that of the prime rate, is skewed to the right. Fifty percent of the observations appear in the third and fourth classes, 4–4.9 % and 5–5.9 %. This distribution is depicted in the histogram shown in Fig. 3.17.

If you were to make a prediction of the treasury bill rate, it would probably be in the 3–6 % range. Again, better methods for predicting observations will be discussed later.

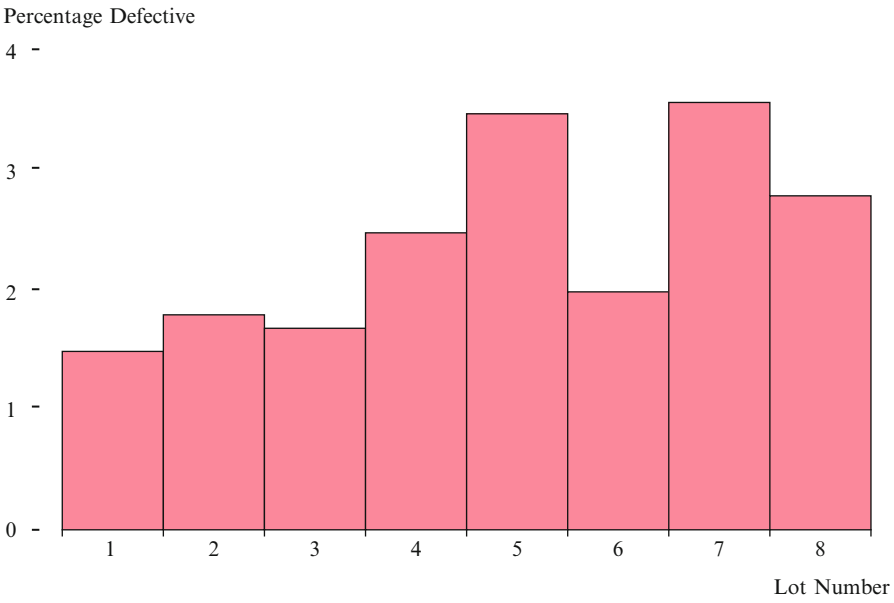


Fig. 3.18 Frequency bar graph of the percentage of defects for each sample lot

Table 3.17 Quality control report on electronic parts

Sample Lot	Sample	Defects	Percentage
1	1,000	15	1.5
2	1,000	20	2.0
3	1,000	17	1.7
4	1,000	25	2.5
5	1,000	35	3.5
6	1,000	20	2.0
7	1,000	36	3.6
8	1,000	28	2.8
Total	8,000	196	2.45 (mean)

3.5.4 Quality Control

Figure 3.18 depicts the quality control data on electronic parts given in Table 3.17. This control chart shows the percentage of defects for each sample lot. Figure 3.18 indicates that both lots 5 and 7 have exceeded the allowed maximum defect level of 3 %. Therefore, the product quality in these two lots should be improved.

3.6 Summary

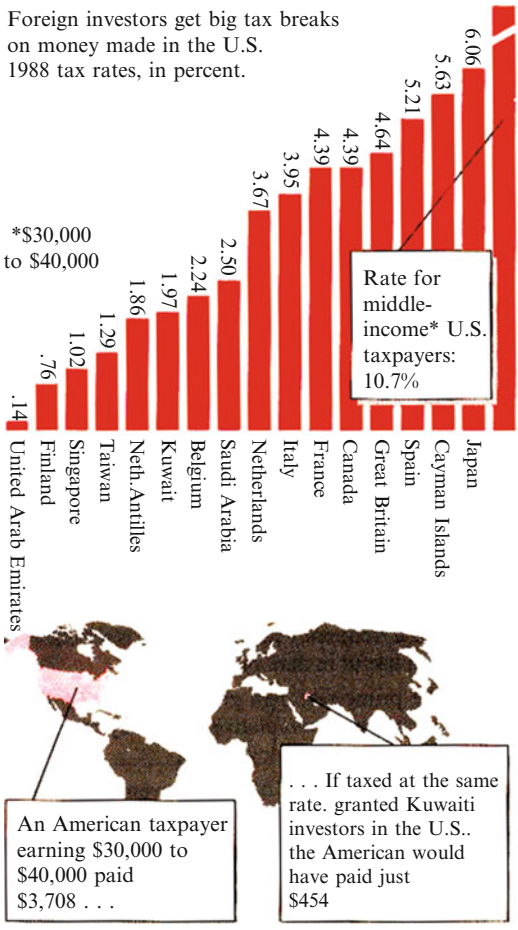
In this chapter, we extended the discussion of Chap. 2 by showing how data can be grouped to make analysis easier. After the data are grouped, frequency tables, histograms, stem-and-leaf displays, and other graphical techniques are used to present them in an effective and memorable way.

Our ultimate goal is to use a sample to make inferences about a population. Unfortunately, neither the tabular nor the graphical approach lends itself to measuring the *reliability* of an inference in data analysis. To do this, we must develop numerical measures for describing data sets. Therefore, in the next chapter, we show how data can be described by the use of descriptive statistics such as the mean, standard deviation, and other summary statistical measures.

Questions and Problems

1. Explain the difference between grouped and nongrouped data.
2. Explain the difference between frequency and relative frequency.
3. Explain the difference between frequency and cumulative frequency.
4. Carefully explain how the concept of cumulative frequency can be used to form the Lorenz curve.
5. Suppose you are interested in constructing a frequency distribution for the heights of 80 students in a class. Describe how you would do this.
6. What is a frequency polygon? Why is a frequency polygon useful in data presentation?
7. Use the prime rate data given in Table 3.6 in the text to construct cumulative frequency and cumulative relative frequency tables.
8. Use the percentage of total revenue spent on advertising listed in Table 3.12 of the text to draw a frequency polygon and a cumulative frequency polygon.
9. On November 17, 1991, the *Home News* of central New Jersey used the bar chart given here to show that foreign investors are taxed at a lower rate than the US citizens.

   (a) Construct a table to show frequency, relative frequency, and cumulative frequency.
   (b) Draw a frequency polygon and a cumulative frequency polygon.

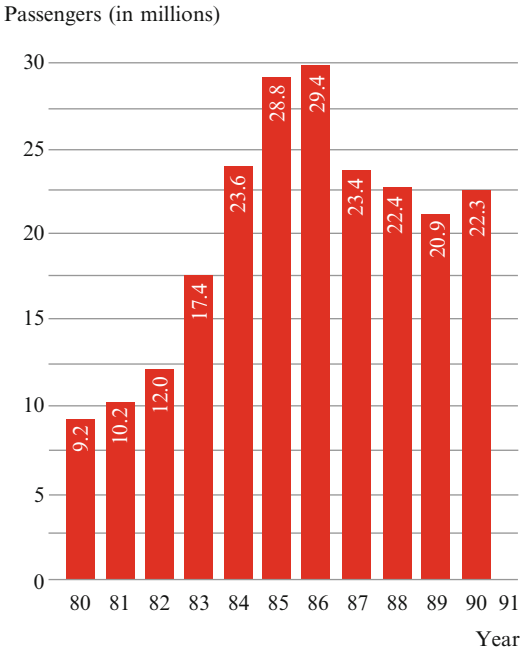


Source: Philadelphia Inquirer, Internal Revenue Service.

Source: *Home News*, November 17, 1991, Reprinted by permission of Knighi-Ridder Tribune News

10. Use the EPS and DPS data given in Table 2.3 in Chap. 2 to construct frequency distributions.
11. Use the data from question 10 to construct a relative frequency graph and a cumulative relative frequency graph for both EPS and DPS.
12. On November 17, 1991, the *Home News* of central New Jersey used the bar chart in the accompanying figure to show the 1980–1991 passenger traffic trends for Newark International Airport.

    (a) Use these data to draw a line chart and interpret your results.
    (b) Use these data to draw a stem-and-leaf diagram and interpret your results.



Source: Port Authority of NY and NJ

Source: *Home News*, November 17, 1991. Reprinted by permission of the publisher

13. An advertising executive is interested in the age distribution of the subscribers to Person magazine. The age distribution is as follows:

Age	Number of subscribers
18–25	10,000
26–35	25,000
36–45	28,000
46–55	19,000
56–65	10,000
Over 65	7,000

(a) Use a frequency distribution graph to present these data.

(b) Use a relative frequency distribution to present these data.

14. Use the data from question 13 to produce a cumulative frequency graph and a cumulative relative frequency graph.

15. Construct stem-and-leaf displays for the 3-month T-bill rate and the prime rate, using the data listed in Table 3.15.

Use the goaltenders' salaries for the 1991 NHL season given in the following table to answer questions 16–20.

Name	Team	Gross salary
Patrick Roy	Montreal Canadiens	$1.056M[a]
Ed Belfour	Chicago Blackhawks	$925,000
Ron Hextall	Philadelphia Flyers	$735,000
Mike Richter	New York Rangers	$700,000
Kelly Hrudey	Los Angeles Kings	$550,000
Mike Liut	Washington Capitals	$525,000
Mike Vernon	Calgary Flames	$500,000
Grant Fuhr	Toronto Maple Leafs	$424,000
John Vanbiesbrouck	New York Rangers	$375,000
Ken Wregget	Philadelphia Flyers	$375,000
Tom Barrasso	Pittsburgh Penguins	$375,000

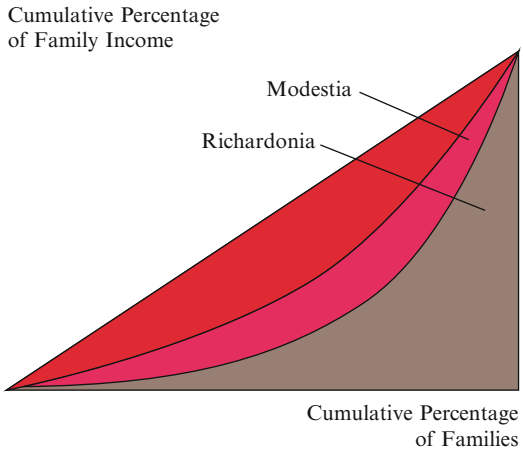
[a]Roy's salary is $500,000 Canadian, and $700,000 Canadian deferred. The salary listed is US equivalent

16. Group the data given in the table into the following groups: $351,000–400,000; 401,000–450,000; 451,000–500,000; 501,000–550,000; 551,000–600,000; 601,000–650,000; 651,000–700,000; over 701,000.
17. Use your results from question 16 to construct a cumulative frequency table.
18. Use your results from question 16 to construct a relative frequency table and a cumulative relative frequency table.
19. Use a bar graph to plot the frequency distribution.
20. Use a bar graph to plot the cumulative relative frequency.
21. Briefly explain why the Lorenz curve shown in Fig. 3.15b has the shape it does.
22. The students in an especially demanding history class earned the following grades on the midterm exam: 86, 75, 92, 98, 71, 55, 63, 82, 94, 90, 80, 62, 62, 65, and 68. Use MINITAB to draw a stem-and-leaf graph of these grades.
23. Use the data given in question 22 to construct a tally table for the grades. Use intervals 51–60, 61–70, 71–80, 81–90, and 91–100.
24. Construct a cumulative frequency table for the tally table you constructed in question 23.
25. Use the data in question 24 to graph the cumulative frequency on a bar chart by using Microsoft Excel.
26. Suppose the Gini coefficient in some country were equal to 0. What would that tell us about income in this country?
27. Suppose the Gini coefficient in another country were equal to 1. What would that tell us about income in this country?

Use the following information to answer questions 28–34. Suppose Weight Watchers has collected the following weight loss data, in pounds, for 30 of its clients.

15, 20, 10, 6, 8, 18, 32, 17, 19, 7, 9, 12, 14, 9, 25, 18, 21, 3, 2, 18, 12, 15, 14, 28, 34, 30, 18, 12, 11, 8

28. Construct a tally table for weight loss. Use 5-lb intervals beginning with 1–5 lb, 6–10 lb, etc.
29. Construct a cumulative frequency table for weight loss.
30. Construct a frequency histogram for weight loss using MINITAB.
31. Construct a frequency polygon for weight loss.
32. Construct a table for the relative frequencies and the cumulative relative frequencies.
33. Graph the relative frequency.
34. Graph the cumulative relative frequency.
35. The following graph shows the Lorenz curves for two countries, Modestia and Richardonia. Which country has the most nearly equal distribution of income?



Use the following information to answer questions 36–41. Suppose a class of high school seniors had the following distribution of SAT scores in English.

SAT score	Number of students
401–450	8
451–500	10
501–550	15
551–600	6
601–650	4
651–700	1

36. Construct a cumulative frequency table.
37. Use a histogram to graph the cumulative frequencies.
38. Construct a frequency polygon.
39. Compute the relative frequencies and the cumulative relative frequencies.
40. Construct a relative frequency histogram.
41. Construct a cumulative relative frequency histogram.

Use the following prices of Swiss stocks to answer questions 42 through 49.

Switzerland (in Swiss francs)	Close	Prev. close
1. Alusuisse	976	982
2. Brown Boveri	3,960	4,080
3. Ciba-Geigy br	3,190	3,240
4. Ciba-Geigy reg	3,080	3,110
5. Ciba-G ptc ctf	3,020	3,040
6. CS Holding	1,920	1,915
7. Hof LaRoch br	8,280	8,300
8. Roce div rt	5,360	5,330
9. Nestle bearer	8,420	8,450
10. Nestle reg	8,310	8,310
11. Nestle ptc ctf	1,570	1,585
12. Sandoz	2,390	2,410
13. Sulzer	465	470
14. Swiss Bank Cp	301	299
15. Swiss Reinsur	2,520	2,530
16. Swissair	667	680
17. UBS	3,230	3,230
18.Winterthur	3,390	3,420
19. Zurich Ins	4,080	4,090

Source: *Wall Street Journal*, November 1, 1991

42. Construct a tally table for the closing stock prices "Close" column. Use 1,000-point intervals beginning with 301–1,300, 1,301–2,301, etc.
43. Compute the change in prices by subtracting the previous closing price from the current closing price.
44. Use your answer to question 43 to construct a tally table. Use 30-point intervals beginning with $-120 \sim -91$,–$90 \sim -61$, etc.
45. Use your answer to question 44 to compute the cumulative frequencies.
46. Use your answer to question 44 to compute the relative and cumulative relative frequencies.
47. Use your answer to question 46 to graph the relative frequency.
48. Use your answer to question 46 to graph the cumulative frequency.
49. Create a frequency polygon using data from question 44.
50. Draw the stem-and-leaf display of DPS of JNJ and Merck during the period 1988–2009 using Table 2.3, in which data on EPS, DPS, and PPS for JNJ, Merck, and S&P 500 during the period 1988–2009 are given.
51. Refer to Table 2.5, in which the balance sheet of JNJ company for the year 2008 and 2009 are given. Draw the pie chart of the composition of the total current asset of JNJ for the year 2008 and 2009, respectively.

    Using Table 2.8, in which the 7 financial ratios of JNJ and Merck during the period 1990–2009 are given.
52. Construct a frequency, cumulative frequency, and relative frequency table for the "price–earnings ratio" (PER) of the JNJ company using class boundaries: $-20.000 < \text{PER} \leq 0.000$, $0.000 < \text{PER} \leq 5.000$, $5.000 < \text{PER} \leq 10.000$, $10.000 < \text{PER} \leq 15.000$, $15.000 < \text{PER} \leq 20.000$, $20.000 < \text{PER} \leq 27.000$.
53. Draw the histogram and frequency polygon of the above frequency distribution.

Chapter 4
Numerical Summary Measures

Chapter Outline

Key Terms

Measure of central tendency	Quartiles
Arithmetic mean	Interquartiles range
Geometric mean	Box and whisker plots
Median	Z score
Mode	Tchebysheff's theorem
Dispersion	Skewness
Variance	Coefficient of skewness
Standard deviation	Pearson coefficient
Mean absolute deviation	Zero skewness coefficient
Range	Positive skewness coefficient
Coefficient of variation	Negative skewness coefficient
Percentile	Kurtosis

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_4,

4.1 Introduction

In this chapter, we extend the graphical descriptive method in data analysis by examining measures of central tendency, dispersion, position, and shape. All these numerical summary measures are important because they enable us to describe a set of data with only a small number of summary statistics. One use of these summary statistics is to compare individual observations from a data set. For example, a student in a statistics class could use one measure of central tendency, the class average, or mean, to determine how well her performance stacks up to the rest of the class. Measures of central tendency can also be used to compare two different sets of data. For example, a statistics teacher interested in comparing the performances of two different statistics classes could take the average, or mean, for each class and compare the two.

We first address four measures *of central tendency*, discussing how they are computed from a data set and how they help us locate the center of a distribution (see Fig. 4.1a). Similarly, we examine measures of *dispersion*, which describe the dispersion, or spread, of a set of observations and therefore of its distribution (see Fig. 4.1b). The coefficient of variation (a measure of relative dispersion) is also investigated. Next, we explore measures of a distribution's *position*. Numerical descriptive measures have also been devised to measure *shape*: the skewness of a distribution (the tendency of a relative frequency distribution to stretch out in one direction or another) and its kurtosis (peakedness). Here, we discuss only the numerical measurement of skewness. The numerical measurement of kurtosis will be discussed in Chap. 9. Finally, we present applications of numerical descriptive statistics in business and economics.

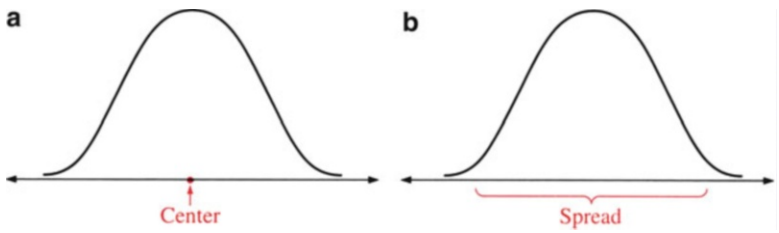


Fig. 4.1 Numerical summary measures: (**a**) central tendency and (**b**) dispersion

4.2 Measures of Central Tendency

The purpose of a *measure of central tendency* is to determine the "center" of a distribution of data values or possibly the "most typical" data value. Measures of central tendency include the arithmetic mean, geometric mean, median, and mode.

Using a quality control example, we will illustrate each of these measures with the following data, which represent the number of defective parts in each of four samples[1]:

5, 8, 14, 3

4.2.1 The Arithmetic Mean

Most of you have calculated your grade point average or average test score in a course by adding all your grade points or scores and dividing by the number of courses or tests. You might not have realized it, but you were calculating the *arithmetic mean*.

The arithmetic mean of a set of raw data is denoted by $x_1, x_2, \ldots, x_N$ (N represents the total number of observations in a population) or $x_1, x_2, \ldots, x_n$ (n represents the sample size). We find it by adding together all the observations and dividing by the number of observations. A sample mean is denoted by $\bar{x}$, a population mean by μ. For the quality control data set, $n = 4$, so

$$\bar{x} = (5 + 8 + 14 + 3)/4 = 30/4 = 7.5$$

Thus, when the observations are $x_1, x_2, \ldots, x_n$, the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.1}$$

The population mean is

$$\mu = \sum_{i=1}^{N} x_i / N \tag{4.2}$$

where N is the total number of observations of the population and the observations are $x_1, x_2, \ldots, x_N$. The summation notation ($\sum$) used in Eqs. 4.1 and 4.2 simply means that the first observation is added to the second and so on, until all the observations have been added.

Example 4.1 Six from Nine to Five. Say we want to find the average annual salary of all secretaries. We believe we can do this on the basis of our knowledge of the annual salaries of six particular secretaries, who each year earn $10,400, $34,000,

[1] Quality control was addressed briefly in Table 3.17 of Chap. 3. This issue will be discussed further in Chaps. 10 and 11.

$14,000, $18,500, $27,000, and $25,800, respectively. This is a sample of $n = 6$, where $x_1 = 10{,}400$, $x_2 = 34{,}000$, $x_3 = 14{,}000$, $x_4 = 18{,}500$, $x_5 = 27{,}000$, and $x_6 = 25{,}800$. We find the sample mean by adding all the observations and dividing by 6:

$$\bar{x} = (x_1 + x_2 + x_3 + x_4 + x_5 + x_6)/6 = 129{,}700/6$$
$$= \$21{,}616.67$$

Our result is a *sample* mean because we are interested in finding the mean annual income of all secretaries on the basis of the annual income of a smaller sample consisting of only six secretaries.

Example 4.2 Arithmetic Average of Stock Rates of Return. As an example of computing the mean of a *population*, suppose an individual owns five stocks that last year returned 15 %, 10 %, −4 %, 7 %, and −10 %. We find the mean of this population by adding all the returns and dividing by $N = 5$. Thus, the population mean is $\mu = (15 + 10 + -4 + 7 + -10)/5 = 18/5 = 3.6$ %.

4.2.2 The Geometric Mean

The *geometric mean* of a set of observations is another measure of central tendency. It can be calculated by multiplying all the observations and taking the product to the l/*N* or the *1*/*n* power, depending on whether the observations come from a finite population or a sample. The sample mean ($\bar{x}_g$) and the population geometric mean (μ_g) can be expressed as follows:

$$\bar{x}_g = (x_1 \cdot x_2 \cdot \cdots \cdot x_n)^{1/n} \tag{4.3}$$

$$\mu_g = (x_1 \cdot x_2 \cdot \cdots \cdot x_N)^{1/N} \tag{4.4}$$

Using the quality control data set, we find that

$$\bar{x}_g = (5 \cdot 8 \cdot 14 \cdot 3)^{1/4} = (1680)^{1/4} = 6.40$$

Note that the geometric mean, 6.4, is smaller than the arithmetic mean, 7.5.

All the observations in Eqs. 4.3 and 4.4 must be positive. It should be noted that the geometric mean is less sensitive to extreme values than is the arithmetic mean. The geometric mean is frequently used in finance to calculate the rate of return on a stock or bond. The reason why the geometric mean is popular in calculating average rates of return is that this method explicitly incorporates the concept of compound

interest (interest received on interest).[2] To avoid negative and zero returns, holding period returns (HPR) are used. An HPR is calculated by taking the rate of return and adding 1. Adding 1 avoids negative numbers and makes it possible to calculate an average return. Now let's use the data given in Example 4.2 to calculate the geometric average of stock rates of return.

Example 4.3 Geometric Average of Stock Rates of Return. Here, we must calculate a geometric mean of the following rates of return: 15 %, 10 %, −4 %, 7 %, and −10 %. To obtain the HPR, we add 1 to each of the returns, which yields 1.15, 1.10, .96, 1.07, and .90. To obtain the geometric mean of the HPR, we multiply the individual HPRs and take the product to the $1/N$ power:

$$\mu_g = [(1.15)(1.10)(.96)(1.07)(.90)]^{1/5} = (1.169)^{1/5} = 1.032$$

To obtain the geometric mean for the conventional rate of return, we subtract 1 from the geometric-mean HPR, arriving at .032 or 3.2 %. Note that the geometric mean is 3.2 %, whereas the arithmetic mean (calculated in Example 4.2) is 3.6 %. In general, the geometric mean is smaller than the arithmetic mean and less sensitive to extreme observations.

4.2.3 The Median

The *median* (Md) is the middle observation of a set of ordered observations if the number of observations is odd; it is the average of the middle pair if the number of observations is even. In other words, if there are N observations, where N is an odd number, the median is the $[(N + 1)/2]$th observation. If N is even, the median is the average of the $(N/2)$th and the $[(N + 2)/2]$th observations. Sometimes the median is a preferred measure of central tendency, particularly when the data include extreme observations that could affect the geometric or arithmetic mean.

Consider again our quality control data. We find the median Md by first constructing an order array:

3, 5, 8, 14

Because N is an even number (4), $\text{Md} = (5 + 8)/2 = 6.5$. The median (6.5) is smaller than the mean (7.5), as indicated in Fig. 4.2. This difference is essentially caused by the extreme value 14. The relationship between mean and median will be discussed in Sect. 4.5.

[2] The advantage of using the geometric average rather than the arithmetic average is discussed by Lee et al. (1990), *Security Analysis and Portfolio Management,* Scott, Foresman, Little, Brown (Chap. 3).

Fig. 4.2 Sample quality control data with mean and median shown

Example 4.4 Median of Stock Rates of Return. Arrange the rate-of-return data in Example 4.2 in numerical order: −10 %, −4 %, 7 %, 10 %, and 15 %. There is an odd number of observations, so the median is the third observation—[(5 + 1)/2] = 3—which in this case is 7 %.

Example 4.5 What Does "Average" Mean? Arithmetic Mean Versus Median for Sample Family Income. The median can be particularly useful when there are a few extreme observations. Consider the following incomes for six sample families: $10,000, $13,400, $15,000, $17,000, $19,000, and $120,000. Although the arithmetic mean of the series is $32,400, the median is ($15,000 + $17,000)/2 = $16,000. The substantial difference between the two means is due mainly to the extreme observation of $120,000. The median is the better measure of central tendency in this example. Note that the median would not change if the fifth and sixth observations were larger. For example, the last number could be $5,000,000 and the median would remain unchanged. Thus using the median is preferred when outlying data could lead to a distorted picture of the mean of a distribution.

Calculations of average income are especially vulnerable to such distortions. Consider the effect of that single $120,000 income if, say, federal assistance for day care were being made available in communities where the average income was under $20,000—and the "average income" of our community of six families were being interpreted as the mean.

Example 4.6 This Teacher Is Really Mean. Students may complain that one or two very high scores raise the class average on an exam and thus lower their letter grade. See Table 4.1, where the "rank on exam" in the third column is obtained by ranking all scores in order from lowest to highest. If the mean is taken as the average score that translates into a grade of C, five of these seven students have scored "below average." Students who see this as unfair are in effect arguing against using the mean exam score as a measure of central tendency. Are they right?

Well, Albert and Sue did score exceptionally high on the exam, and they do in fact raise the mean score dramatically. Should the teacher base the class grades on the mean of 72.43 or use some other measure of central tendency? The median (here it is 62), which lies in the middle and is not altered by the extreme values that affect the mean, may be a better measure of central tendency in this case. (Juan and Mary, whose grades have just risen from D to B and C, respectively, will certainly think so.)

Table 4.1 Student exam scores

Student	Score	Rank on exam
Kim	60	2
Mary	62	4
Tom	55	1
Ann	61	3
Juan	70	5
Albert	99	6
Sue	100	7
Total	507	

Mean $= \bar{x} = 72.43$
Median $= 62$

Table 4.2 Sales of personal computers

Type	Number sold
IBM PS2/M30	487
IBM PS2/M50	201
IBM PS2/M70	432
Compaq	506

4.2.4 The Mode[3]

The *mode* of a set of observations is the value that occurs the most times. In cases of a tie, it may assume more than one value. The mode is most useful when we are dealing with data that are in categories where the mean and median are not useful. For example, suppose that a computer sales representative sells the brands and numbers of computers shown in Table 4.2. Here, it makes no sense to take the mean or median of the data, because the categories are mutually exclusive. Instead, the sales rep is interested in the most popular and the least popular products. Thus, he or she wants to know which is the *modal* class (it is the Compaq computer) because that class contains the highest number of computers sold.

The main disadvantage of the mode is that it does not take the nonmodal observations into consideration. Thus, in the computer example, the mode does not reflect the facts that the IBM M30 has almost as many sales as the Compaq model and that the IBM M70 has almost the same amount of sales as the IBM M30. As another example, suppose a sample of the incomes of workers is taken and the arithmetic mean is $25,746. Suppose further that the observation $38,500 appears the most times and therefore is the mode. Obviously, the mode is not a good measure of central tendency here, because it is so far away from the mean. This problem occurs often, and researchers must be aware of the limitations of this and other statistical measures.

Example 4.7 The Model Wears 4, but the Modal Is 7. Suppose a clerk in a shoe store sells eight pairs of shoes in the following sizes: 5, 7, 7, 7, 4, 5, 10, and 11.

[3] Relationships among mean, median, and mode will be discussed in Sect. 4.5.

The modal shoe size is 7 because it appears the greatest number of times. If this result is obtained regularly, it is certainly something the purchasing manager wants to know. Although the mean and median are more widely used as measures of central tendency in business and economics, the mode gives useful information on the most numerous value in a set of observations.

The numerical example discussed in Example 4.7 is a unimodal distribution. The following is a bimodal distribution: 5, 5, 7, 7, 7, 8, 9, 10, 10, 10 (7 and 10 are the modes). It should be noted that if each different number has the same frequency, there is no mode. An example of a case of no mode is 1, 1, 2, 2, 4, 4.

4.3 Measures of Dispersion

The mean, median, and mode all give us information about the central tendency of a set of observations, but these measures shed no light on the *dispersion*, or spread, of the data. For example, suppose a professor gives a test to two classes and the mean for each class is 75. However, suppose that all the students in the first class scored in the 70s, with a high of 79 and a low of 70. In the second class, the lowest score was 42 and the highest 97. It is obvious that the scores in the second class are more widely dispersed, or spread, around the mean than the scores in the first. In this section, we discuss measures of dispersion: the variance, standard deviation, mean absolute deviation, range, and coefficient of variation. We will use our now-familiar quality control data (3, 5, 8, 14) to illustrate these different measures.

4.3.1 The Variance and the Standard Deviation

Suppose we have a set of observations from a population $x_1, x_2, \ldots, x_N$. We are interested in finding a dispersion measure, so it would seem natural to calculate the deviations from the mean $(x_1 - \mu), (x_2 - \mu), \ldots, (x_N - \mu)$. But the negative deviations from the mean cancel out the positive deviations,[4] so the sum of these deviations will always be zero, which sheds no light on the extent of dispersion. To avoid this problem, we square and sum the deviations (distances) to give an indication of the total dispersion:

$$(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2$$

If we take this sum and divide by the number of observations, N, we arrive at the *variance,* which represents the average squared deviation (distance) from the mean. The *population variance* is denoted by σ^2, as indicated in Eq. 4.5, and the *sample*

[4] Because $(x_1 - \mu) + (x_2 - \mu) + \cdots + (x_N - \mu) = \sum_{i=1}^{N} x_i - N_\mu = N_\mu - N_\mu = 0.$

variance by s^2, as indicated in Eq. 4.7. The *standard deviation* is the square root of the variance; it is denoted by σ for the population (Eq. 4.6) and by s for the sample (Eq. 4.8).

Population variance	Population standard deviation
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$ (4.5)	$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$ (4.6)
Sample variance	Sample standard deviation
$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ (4.7)	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ (4.8)

A shortcut formula that can be used to compute the variance and standard deviation for samples and populations is given in Appendix 1.

Using the quality control data, we calculate the sample variance and standard deviation in accordance with Eqs. 4.7 and 4.8. Recall that $\bar{x} = 7.5$.

	x	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	3	−4.5	20.25
	5	−2.5	6.25
	8	.5	.25
	14	6.5	42.25
Total	30	$\sum(x_i - \bar{x}) = 0$	$\sum(x_i - \bar{x})^2 = 69$

Substituting $\sum(x - \bar{x})^2 = 69$ into Eqs. 4.7 and 4.8, we obtain

$$s^2 = 69/(4-1) = 23$$
$$s = 4.80$$

Note that we use the divisor $(n - 1)$ instead of n to calculate the sample variance. This is because using the divisor $(n - 1)$ yields a more precise estimate of σ^2 than dividing the sum of squared distances by n.[5]

For purposes of comparison, let's calculate the variance and the standard deviation of another set of quality control data (3, 4, 5, 6):

$$\bar{x} = (3 + 4 + 5 + 6)/4 = 4.5$$

$$s^2 = \frac{(3-4.5)^2 + (4-4.5)^2 + (5-4.5)^2 + (6-4.5)^2}{4-1} = 1.67$$
$$s = 1.29$$

[5] "More precise" means that s^2 with a divisor of $(n - 1)$ instead of n has the mean σ^2. See Sect. 9.4 for the proof.

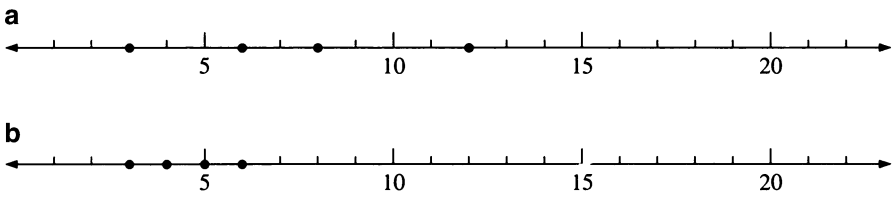


Fig. 4.3 Two sets of quality control data: (**a**) first set of data and (**b**) second set of data

Table 4.3 Worksheet for data on net profit ratios

Net profit margin (x)	$(x - \bar{x})$	$(x - \bar{x})^2$	x^2
5.6	1.778	3.16	31.36
2.7	−1.122	1.26	7.29
7.3	3.478	12.10	53.29
3.5	−.322	.103	12.25
.01	−3.812	14.53	.00
19.11	0	31.1533	104.19

Mean $= \bar{x} = 19.11/5 = 3.822$
Variance $= s^2 = 31.153/4 = 7.79$
Standard deviation $= s = \sqrt{7.79} = 2.79$

We see, then, that the variance of the second set of quality control data is smaller than the variance of the first. The smaller variance of the second set of data is graphically represented in Fig. 4.3.

Example 4.8 Variability of Profit Margin. Suppose we want to calculate the variance and standard deviation for the net profit margins indicated, for a certain firm over a 5-year period, in Table 4.3. Because this is a sample, $n = 5$.

The next computations show how to calculate the variance and standard deviation by using the shortcut formulas of Eqs. 4.7a and 4.8a, as indicated in Appendix 1:

$$\text{Variance} = s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

$$= \frac{104.19 - 5(3.822)^2}{4} = 7.79$$

$$\text{Standard deviation} = s = \sqrt{7.79} = 2.79$$

Note that the answers are the same as those derived with the standard formulas as indicated in Table 4.3.

Table 4.4 Worksheet for sales data

Sales (x)	$x - \bar{x}$	$(x - \bar{x})^2$	x^2
2.3	−.425	.181	5.29
1.1	−1.625	2.64	1.21
.7	−2.025	4.100	.49
6.8	4.075	16.60	46.24
10.9	0	23.53	53.23

Example 4.9 Variability of Sales. Suppose a sample of sales is taken from four firms and that the figures are $2.3 million, $1.1 million, $.7 million, and $6.8 million (see Table 4.4). Substituting related information into Eqs. 4.1, 4.7, and 4.8, we obtain

$$\text{Mean} = \bar{x} = 10.9/4 = 2.725$$

$$\text{Variance} = s^2 = 23.53/(4-1) = 7.8$$

$$\text{Standard deviation} = s = \sqrt{7.8} = 2.8$$

Calculating the variance and standard deviation via the shortcut formula of Eqs. 4.7a and 4.8a, we get

$$s^2 = \frac{53.23 - 4(2.725)^2}{3} = 7.8$$

$$s = \sqrt{7.8} = 2.8$$

Again, the results are identical to those obtained with the standard formula.

4.3.2 The Mean Absolute Deviation

Rather than squaring the deviations from the mean, we can arrive at another useful measure by calculating the absolute deviations from the mean or median and then dividing by the number of observations to obtain the average absolute deviation from the mean or median. This measure, called the *mean absolute deviation* (MAD), is defined as follows:

$$\text{MAD} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \quad \text{or} \quad \frac{\sum_{i=1}^{n} |x_i - \text{Md}_s|}{n} \tag{4.9}$$

where n is the sample size, $\bar{x}$ is the sample mean, and Md_s is the sample median.

If population data instead of sample data are used, then

$$\text{MAD} = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N} \quad \text{or} \quad \frac{\sum_{i=1}^{N} |x_i - \text{MAD}_p|}{N} \tag{4.10}$$

where N is the total observations of population, μ is the population mean, and MAD_p is the population median. Let us calculate $|x_i - \bar{x}|$ and $|x_i - \text{MD}_s|$ for our quality control data. Recall that $\bar{x} = 7.5$ and $\text{Md}_s = 6.5$.

	x_i	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i - \text{Md}_s\rvert$
	3	4.5	3.5
	5	2.5	1.5
	8	.5	1.5
	14	6.5	7.5
Total	30	14	14

Substituting $\sum |x_i - \bar{x}| = 14$ and $\sum |x_i - \text{Md}_s| = 14$ into Eq. 4.9, we obtain

$$\text{MAD} = 14/4 = 3.5 \quad \text{if sample mean is used}$$

$$\text{MAD} = 14/4 = 3.5 \quad \text{if sample median is used}$$

One advantage of this measure is that it is not influenced so much as the variance by extreme observations. A second advantage is that the MAD is easier to interpret than the standard deviation. It is much easier to form a mental picture of the average deviation from the mean than to visualize the square root of the squared deviation from the mean! The MAD is not used much in statistical analysis, however, because complications can arise from its use in making inferences about a population on the basis of sample observations alone.

Example 4.10 Variability of Inflation Forecast. Assume that a population of inflation forecasts for next year consists of the following values: 7 %, 5 %, 4 %, 2 %, and 1 %. The worksheet for calculating MAD by using Eq. 4.10 is given in Table 4.5.

The MAD we find by using the mean is 1.84, and the MAD we find by using the median is 1.80. In this case, the results are not identical because the distribution is not symmetric. If the distribution were highly skewed (like that of the student exam scores given in Table 4.1), the MAD found in terms of the mean would be very different from that found in terms of the median. Should we use the mean or the median, then, in calculating the mean absolute deviation? That depends on which measure of central tendency we believe is best for describing our distribution.

Table 4.5 Inflation forecasts

Forecast (x, %)	$x - \mu = x - 3.8 - 3.8$	$\lvert x - \mu \rvert$	$\lvert x - \text{median} \rvert$
7	3.2	3.2	3
5	1.2	1.2	1
4	.2	.2	0
2	−1.8	1.8	2
1	−2.8	2.8	3
		9.2	9

$\mu = 19/5 = 3.8$. $\text{Md}_p = 4$
MAD = 9.2/5 = 1.84
MAD = 9/5 = 1.80

4.3.3 The Range

The range is one of the easiest measures of dispersion to calculate and interpret. The *range* is simply the difference between the highest and lowest values:

$$R = x_{\max} - x_{\min} \tag{4.11}$$

where R = range, $x_{\max}$ = the largest value of all observations, and $x_{\min}$ = the smallest value of all observations. The range of our quality control data is

$$R = 14 - 3 = 11$$

The disadvantage of using this measure is that it takes into consideration only these two values. Thus, it is easily thrown off by extreme values. In contrast, the variance, standard deviation, and MAD use all the observations. Despite this problem, the range has some value, for example, the typical range of temperatures in New England during the winter tells us a lot about that area's climate. The *Wall Street Journal* and other newspapers use the range when they report the 52-week high and low for each stock price per share. For example, on January 9, 1991, the 52-week high and low for IBM and Digital Equipment Corporation were $\$123\frac{1}{8}$ −$95 and $\$95\frac{1}{8} - \$45\frac{1}{2}$, respectively (see Fig. 4.4). Comparing these two ranges reveals that the price range of Digital Equipment ($\$49\frac{5}{8}$) was much greater than the price range of IBM ($\$28\frac{1}{8}$).

4.3.4 The Coefficient of Variation

The *coefficient of variation* (CV) is the ratio of the standard deviation to the mean. The coefficient of variation for sample data can be defined as

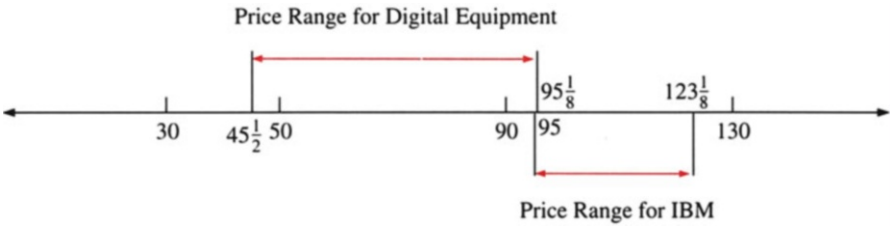


Fig. 4.4 Ranges of stock prices per share for Digital Equipment and IBM (in dollars)

$$\mathrm{CV}_x = s/\bar{x} \tag{4.12}$$

For our quality control data,

$$\mathrm{CV}_x = 4.8/7.5 = .64$$

The coefficient of variation is particularly useful when we must compare the variabilities of data sets that are measured in different units. For example, suppose a researcher wants to see how Japanese and American wage incomes compare in variability. Because workers are paid in yen in Japan and in dollars in the United States, it would be difficult to make this comparison using, say, the standard deviation. Being a *relative* expression—that is, a ratio—the coefficient of variation neatly avoids this problem.

The coefficient of variation is also useful when we are comparing data of the same type from different time periods. Suppose, for example, that the mean sales for a firm from 1980 to 1985 were \$.5 million with a standard deviation of \$50,000. The mean sales for the same firm from 1986 to 1991 were \$3.2 million with a standard deviation of \$100,000. If we compared the standard deviations and concluded that sales in the 1980s were less variable, we would obtain a distorted picture. The value of sales during the earlier period was much lower than during the later period, so the resulting standard deviation is almost certain to be smaller. Therefore, we use the coefficient of variation, which we calculate as .1 for the earlier period and .03 for the later. Sales were actually less variable from 1986 to 1991.

Example 4.11 Using Coefficient of Variation to Analyze the Volatility of Stocks. Whenever we compare two different stocks (A and B), it is useful to know two particular statistics: (1) the average, or mean, of the stocks' rates of return and (2) the standard deviation of the returns, which is an indicator of risk. A high rate of return and low risk are desirable, so having a high mean and a low standard deviation is the most desirable combination. Let's say the average rates of return ($\bar{R}$) and standard deviations (σ) for these two stocks are

	$\bar{R}$	σ
Stock A	10 %	1 %
Stock B	12 %	1.5 %

Stock B has a higher mean return ($\bar{R}$), but it also has a larger standard deviation (σ), which means it is more risky. Because A is less risky but B has a higher expected return, the choice between the two is not obvious. By using the coefficient of variation, however, we can find the amount of risk (standard deviation) per unit of expected return, which is an appropriate measure of relative variability that combines risk and rate of return. Substituting the values of $\bar{R}$ and σ into Eq. 4.12 yields

$$\mathrm{CV}_A = .01/.10 = .10$$

$$\mathrm{CV}_B = .015/.12 = .125$$

Thus, stock B has a greater risk per unit of expected return than stock A.

4.4 Measures of Relative Position

In some situations, we may want to describe the relative position of a particular measurement in a set of data. In this section, we discuss three measures of relative standing: percentiles, quartiles, and Z scores.

To illustrate these measures, suppose the personnel managers of Johnson & Johnson have administered an aptitude test to 40 job applicants. Their scores are presented in Table 4.6. The mean of the data is $\bar{x} = 58.45$, and the standard deviation is $s = 22.99$. The sample size is $n = 40$.

4.4.1 Percentiles, Quartiles, and Interquartile Range

One useful way of describing the relative standing of a value in a set of data is through the use of percentiles. *Percentiles* give valuable information about the rank of an observation. Most of you are familiar with percentiles from taking standardized college admissions tests such as the SAT or ACT. These tests assign each student not only a raw score but also a percentile to indicate his or her relative performance. For example, a student scoring in the 85th percentile scored higher than 85 % of the students who took the test and lower than $(100 - 85) = 15$ % of those who took it.

Let $x_1, x_2, \ldots$, be a set of measurements arranged in ascending (or descending) order. The Pth percentile is a number x such that P percent of the measurement fall below the Pth percentile and $(100 - P)$ percent fall above it.

Quartiles are merely particular percentiles that divide the data into quarters. The 25th percentile is known as the first quartile (Q_1), the 50th percentile is the second (Q_2), and the 75th percentile is the third (Q_3).

Table 4.6 *Ordered* array of aptitude test scores for 40 job applicants ($\bar{x} = 58.45$, $s = 22.99$)

i	x	i	x	i	x	i	x
1.	20	11.	42	21.	56	31.	78
2.	21	12.	43	22.	58	32.	80
3.	23	13.	43	23.	59	33.	81
4.	25	14.	46	24.	61	34.	85
5.	30	15.	48	25.	62	35.	90
6.	35	16.	50	26.	65	36.	92
7.	36	17.	51	27.	68	37.	96
8.	39	18.	52	28.	70	38.	98
9.	40	19.	54	29.	71	39.	99
10.	41	20.	55	30.	75	40.	100

To approximate the quartiles from a population containing N observations, the following positioning point formulas are used:

$$Q_1 = \text{value corresponding to the } \frac{N+1}{4} \text{ ordered observation}$$

$$Q_2 = \text{median, the value corresponding to the } \frac{2(N+1)}{4} = \frac{N+1}{2} \text{ ordered observation}$$

$$Q_3 = \text{value corresponding to the } \frac{3(N+1)}{4} \text{ ordered observation}$$

The formulas given for Q_1 and Q_3 sometimes are defined as the $(N + 1)/4$th and $(3N + 1)/4$th observations, respectively. If Q_1, Q_2, or Q_3 is not an integer, then the interpolation method can he used to estimate the value of the corresponding observation.

Interquartiles range (IQR), a measure commonly used in conjunction with quartiles, can be defined as

$$\text{IQR} = Q_3 - Q_1 \tag{4.13}$$

The interquartile range has an easy and sometimes convenient interpretation. For large data sets, it is the range that contains the middle half of all the observations.

Now we use the Johnson & Johnson applicant data to determine the first quartile (Q_1), second quartile (Q_2), third quartile (Q_3), and interquartiles range (IQR). First we find the locations $Q_1 = 41(.25) = 10.25$, $Q_2 = 41(.5) = 20.5$, and $Q_3 = 41(.75) = 30.75$. On the basis of these locations and the information in Table 4.6, we find that the scores for Q_1, Q_2, and Q_3 are 41.25, 55.5, and 77.25. Then, according to Eq. 4.13, IQR $= 77.25 - 41.25 = 36$.

Example 4.12 Finding One Applicant's Percentile. If James Fleetdeer received an aptitude test score of 92, what is the percentile value?

Because Table 4.6 is arranged in ascending order, Mr. Fleetdeer's 5th-highest score is the 36th-smallest value (out of a total of 40). Hence, the percentile is

$$P = \frac{36}{40}.100 = 90$$

4.4.2 Box and Whisker Plots: Graphical Descriptions Based on Quartiles

A *box and whisker plot* is a graphical representation of a set of sample data that illustrates the lowest data value (L), the first quartile (Q_1), the median (Q_2, Md), the third quartile (Q_3), the interquartile range (IQR), and the highest data value (H).

In the last section, the following values were determined for the aptitude test scores in Table 4.6: $L = 20$, $Q_1 = 41 + .25(42 - 41) = 41.25$, $Q_2 = \text{Md} = 55.5$, $Q_3 = 75 + .75(78 - 75) = 77.25$, $IQR = 36$, and $H = 100$.

A box and whisker plot of these values is shown in Fig. 4.5. The ends of the box are located at the first and third quartiles, and a vertical bar is inserted at the median. Consequently, the length of the box is the interquartile range. The dotted lines are the whiskers; they connect the highest and lowest data values to the end of the box. This means that approximately 25 % of the data values will lie in each whisker and in each portion of the box. If the data are symmetric, the median bar should be located at the center of the box. Consequently, the location of the bar informs us about any skewness of the data; if the bar is located in the left (or right) half of the box, the data are skewed right (or left), as defined in the next section.

In Fig. 4.5, the distribution of the data is skewed to the right because the median bar is located in the left. A box and whisker plot using MINITAB is shown in Fig. 4.6. In this figure, a rectangle (the box) is drawn with the ends (the hinges) drawn at the first and third quartiles (Q_1 and Q_3). The median of the data is shown in the box by the symbol +. There are two boxes in Fig. 4.6. The only difference is that the second specifies the starting value at 15.

Example 4.13 Using MINITAB to Compute Some Important Statistics of 40 Aptitude Test Scores. The MINITAB/PC input and printout are presented in Fig. 4.7. This printout presents mean, median, standard deviation, L (MIN), Q_1, Q_3, and H (MAX), which we have calculated and analyzed before. Note that the MINITAB/PC can calculate this information very effectively. In Fig. 4.7, 40 aptitude test scores are first entered into the PC. Then ten statistics will automatically print if the command "MTB > describe Cl" is entered. Two of those statistics, TRMEAN and SEMEAN, are not discussed in this book.

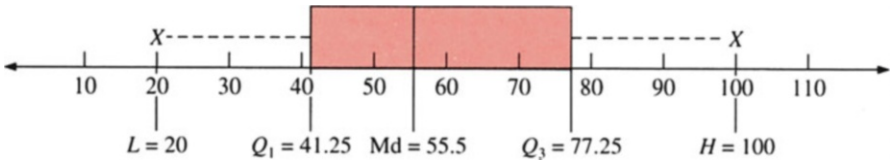


Fig. 4.5 Box and whisker plot for 40 aptitude test scores (Data in Table 4.6)

```
MTB > NAME C1 'TEST'
MTB > SET INTO 'TEST'
DATA> 20 21 23 25 30 35 36 39 40 41 42 43 43 46 48 50 51 52 54 55 56 58 59 61
DATA> 62 65 68 70 71 75 78 80 81 85 90 92 96 98 99 100
DATA> END
MTB > GSTD
* NOTE  * Standard Graphics are enabled.
          Professional Graphics are disabled.
          Use the GPRO command to enable Professional Graphics.
MTB > BOXPLOT USING 'TEST'
```

Fig. 4.6 Box and whisker plot of aptitude test scores using MINITAB

```
MTB > NAME C1 'TEST'
MTB > SET INTO 'TEST'
DATA> 20 21 23 25 30 35 36 39 40 41 42 43 43 46 48 50 51 52 54 55 56 58 59 61
DATA> 62 65 68 70 71 75 78 80 81 85 90 92 96 98 99 100
DATA> END
MTB > DESCRIBE 'TEST'

Descriptive Statistics

Variable      N      Mean    Median   Tr Mean     StDev   SE Mean
TEST         40     58.45     55.50     58.28     22.99      3.63

Variable    Min       Max        Q1        Q3
TEST      20.00    100.00     41.25     77.25
```

Fig. 4.7 The MINITAB/PC input and printout of some important statistics of 40 aptitude test scores

4.4.3 Z Scores

A sample *Z score*, which is based on the mean $\bar{x}$ and standard deviation s of a data set, is defined as

$$Z = \frac{x - \bar{x}}{s} \tag{4.14}$$

Like a percentile, a Z score expresses the relative position of any particular data value in terms of the number of standard deviations above or below the mean. Recall from Example 4.12 that Mr. Fleetdeer had a score of 92 on the test. For this score, $\bar{x} = 58.45$ and $s = 22.99$, as indicated in Table 4.6. His score of 92 is in the 90th percentile. The corresponding Z score is

$$Z = \frac{92 - 58.45}{22.99} = 1.46$$

This means that Mr. Fleetdeer's score of 92 is 1.46 standard deviations to the *right* of (above) the mean. Thus, if Z is positive, it indicates how many standard deviations x is *above* the mean.

A negative value implies that x is to the left of (below) the mean. Look at Table 4.6 again. What is the Z score for the person who got a score of 30 on Johnson & Johnson's aptitude examination?

$$Z = \frac{30 - 58.45}{22.99} = -1.24$$

This individual's score is 1.24 standard deviations *below* the mean.

As a rule of thumb, for mound-shaped data sets, approximately 68 % of the observations have a Z score between -1 and 1 and approximately 95 % of the observations have a Z score between -2 and 2.[6]

Z scores of the aptitude test scores indicated in Table 4.6 are calculated and listed in Table 4.7. From Table 4.7, we find that 67.5 % (27/40) of these observations have a Z score between -1 and 1. All of the observations have Z scores between -2 and 2.

4.5 Measures of Shape

A basic question in many applications is whether data exhibit a symmetric pattern. Skewness and kurtosis are two important characteristics that determine the shape of a distribution.

4.5.1 Skewness

In addition to measures of central tendency and dispersion, there are measures that give information on the skewness of the distribution. The *skewness* indicates whether the distribution is skewed to the left or right in relation to the mean or is symmetric about the mean. The population skewness for *raw* data is given by

$$\mu_3 = \frac{\sum_{i=1}^{N} (x_i - \mu)^3}{N} \tag{4.15}$$

[6] Z scores and their application will be explored further in Chap. 7. This rule of thumb is derived from *Tchebysheff's theorem*, defined as follows: in any data set, the proportion of items within $\pm k$ standardized deviations of the mean is at least $1 - (1/k)^2$, where k is any number greater than 1.0.

Table 4.7 Z scores of aptitude test scores for 40 applicants

i	x_i	Z_i	i	x_i	Z_i
1	20	−1.6728	21	56	−.1066
2	21	−1.6293	22	58	−.0196
3	23	−1.5422	23	59	.0239
4	25	−1.4552	24	61	.1109
5	30	−1.2377	25	62	.1544
6	35	−1.0202	26	65	.2850
7	36	−.9767	27	68	.4155
8	39	−.8462	28	70	.5025
9	40	−.8027	29	71	.5460
10	41	−.7592	30	75	.7200
11	42	−.7157	31	78	.8505
12	43	−.6721	32	80	.9375
13	43	−.6721	33	81	.9810
14	46	−.5416	34	85	1.1551
15	48	−.4546	35	90	1.3726
16	50	−.3676	36	92	1.4596
17	51	−.3241	37	96	1.6336
18	52	−.2806	38	98	1.7206
19	54	−.1936	39	99	1.7641
20	55	−.1501	40	100	1.8076

We can scale the result by dividing μ_3 by σ^3. This gives us the *coefficient of skewness* (CS):

$$\text{CS} = \frac{\mu_3}{\sigma^3} \tag{4.16}$$

The estimate of μ_3 for a sample can be defined as

$$\text{Skewness} = \sum_{i=1}^{n} (x_i - \bar{x})^3 / n \tag{4.15a}$$

The sample coefficient of skewness (SCS) can be defined as

$$\text{SCS} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3 / n}{s^3} \tag{4.16a}$$

An alternative measure of skewness is given by the *Pearson coefficient*, which is defined as

$$\text{Pearson coefficient} = 3(\text{mean} - \text{median})/(\text{standard deviation}) \tag{4.16b}$$

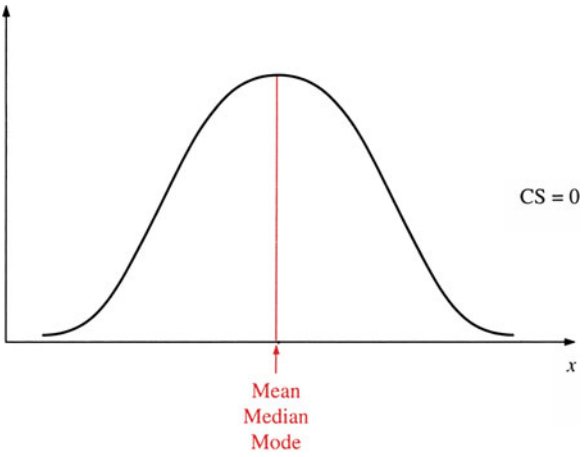


Fig. 4.8 Symmetric distribution

Returning to our quality control data, we can calculate the skewness as follows (recall that $\bar{x} = 7.5$ and $s = 4.8$).

	x	$(x - \bar{x})^3$
	3	−91.13
	5	−15.63
	8	.13
	14	274.63
Total	30	168

Substituting $\sum(x_i - \bar{x})^3 = 168$, $n = 4$, and $s = 4.8$ into Eqs. 4.15a and 4.16a, we obtain the skewness and the sample coefficient of skewness as

$$\text{Skewness} = \frac{168}{4} = 42$$

$$\text{SCS} = \frac{42}{(4.8)^3} = .38$$

This implies that the quality control data are skewed to the right.

A *zero skewness coefficient* means that the distribution is symmetric with mean = median (see Fig. 4.8), which is also equal to the mode if the distribution is unimodal.

A *positive skewness coefficient* means that the distribution is skewed to the right, or positively skewed, and that the mode (most observations) and median lie below the mean (see Fig. 4.9). A *negative skewness coefficient* means that the distribution is skewed to the left, or negatively skewed, and that the mode and median lie above the mean (see Fig. 4.10).

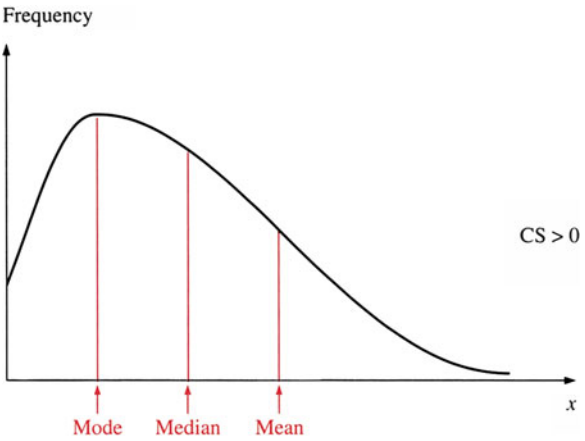


Fig. 4.9 Positively skewed distribution

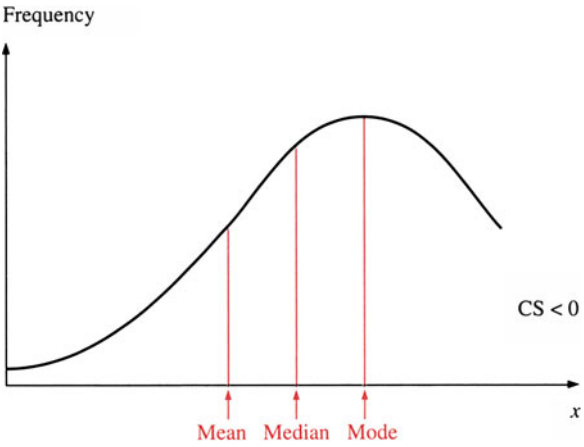


Fig. 4.10 Negatively skewed distribution

4.5.2 Kurtosis

Skewness reflects the tendency of a distribution to be stretched out in a particular direction. Another measure of shape, referred to as *kurtosis*, measures the peakedness of a distribution. In principle, the kurtosis value is small if the frequency of observations close to the mean is high and the frequency of observations far from the mean is low. Because kurtosis is not so frequently used as other numerical summary measures, it is not pursued in further detail here. However, the numerical calculation of kurtosis will be discussed in Chap. 9. Both skewness and kurtosis measures are useful in analyzing stock rates of return (see Chap. 9).

4.6 Calculating Certain Summary Measures from Grouped Data (Optional)

In this section, we discuss how to calculate mean (arithmetic average), median, mode, variance, standard deviation, and percentiles for grouped data.

4.6.1 The Mean

Sometimes data are in grouped form, so calculating the mean from raw data is impossible. Recall from Chap. 3 that raw data are sometimes grouped into classes. For example, a teacher might group exam scores into A's (90–100), B's (80–89), and so on. In cases such as this, we can estimate the mean from the grouped data by multiplying the midpoint of each class by the number of observations and dividing by the total number of observations (N):

$$\mu = \sum_{i=1}^{k} f_i m_i / N \tag{4.17}$$

where f_i = the frequency or number of observations in the ith group, m_i = the midpoint of the ith group, and k = the number of groups. Note that

$$\sum_{i=1}^{k} f_i = N$$

Although this is the formula for estimating the population mean from grouped data, we estimate the sample mean in the same manner by substituting n for N.

Example 4.14 Finding the Mean of Market Rates of Return in Terms of Grouped Data. Table 4.8 presents a frequency distribution for the rate of return on the S&P composite stock index from 1990 to 2009; it has eight classes or groups. Suppose we do not have access to the raw data that underlie this frequency distribution (which, however, are shown in Table 4.9). Will our calculated group mean be reasonably close to the true mean? In this example,

$$\sum_{i=1}^{7} f_i = 19.$$

Following Eq. 4.17, we calculate the group mean as

Table 4.8 Frequency distribution of annual market rates of return

Class	Midpoint (m_i)	Class frequency (f_i)	$m_i f_i$
−.40 to −.30	−0.385	1	−0.385
−.29 to −.20	−0.234	1	−0.234
−.19 to −.10	−0.116	2	−0.232
−.09 to .00	−0.040	2	−0.081
.01 to .10	0.054	5	0.270
.11 to .20	0.166	2	0.331
.21 to .30	0.257	4	1.028
.30 to .40	0.326	2	0.651

$\sum_{i=1}^{n} m_i f_i = 1.35$

Table 4.9 Annual market rates of return in terms of S&P 500 (1990–2009)

Year	Rate of return
90	−0.0656
91	0.2631
92	0.0446
93	0.0706
94	−0.0154
95	0.3411
96	0.2026
97	0.3101
98	0.2667
99	0.1953
00	−0.1014
01	−0.1304
02	−0.2337
03	0.2638
04	0.0899
05	0.0300
06	0.1362
07	0.0353
08	−0.3849
09	0.2345
Sum	1.5525

$$\sum_{i=1}^{7} f_i m_i / 19 = 1.35/19 = .0711.$$

The actual mean of the raw data (see Table 4.9) is 1.5525/19 = .0817. The outcome of our test suggests that the mean of the grouped data is a fairly accurate measure of the true mean of the series.

4.6.2 The Median

Although a median can also be calculated for grouped data, it may be impossible to determine the exact value of the median if individual data values are not available. However, we can approximate the median value by first assuming that data fall equally throughout the median class. For example, suppose we have five students who scored between 90 and 100 on the exam. By assuming that the observations are equally spaced, we can hazard an educated guess of the five students' scores. Because the width of this class is 10 and because there are five students in the class, an assumption of equal spacing means each pair of "adjacent" scores should be separated by 2 points. Making this assumption for all the classes enables us to find what is *approximately* the median or middle score.

To obtain the median for grouped data, find the class in which the median observation appears. Then apply the following formula to estimate the median:

$$m = L + \frac{(N/2 - F)}{f}(U - L) \tag{4.18}$$

where L and U are the lower and upper boundaries, respectively, of the class that contains the median; f is the frequency in this class; and F is the cumulative frequency of the observations in the classes prior to this class.

Example 4.15 Finding the Median of Stock Rates of Return in Terms of Grouped Data. Referring to the grouped data in Example 4.14, we know that the median is the 11th observation. Thus, we know that the median is in the .110 to .200 class. F is equal to 9 because nine observations occurred before the class; f is equal to 8 because there are eight observations in the class; and the lower (L) and upper (U) boundaries of the class are110 and .200, respectively. Hence, the median estimate is

$$m = .110 + \frac{(21/2 - 9)}{8}(.200 - .110) = .1269$$

The median for the raw data is .1240, which is similar to that calculated from grouped data.

4.6.3 The Mode

For nongrouped data, the mode of a set of observations is the value that occurs the most times; for grouped data, the modal class is the one with the highest frequency. Like nongrouped data, grouped data can have more than one class as modal classes.

In Example 4.14 on market rates of return, the modal class, .11 to .20, contains eight observations.

4.6.4 Variance and Standard Deviation

Note that both of the variance formulas yield the same answer.

We can calculate the standard deviation and variance for *grouped* data by using the following formulas:

Population variance	Population standard deviation
$\sigma^2 = \frac{\sum_{i=1}^{k} f_i(m_i-\mu)^2}{N}$ (4.19)	$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i(m_i-\mu)^2}{N}}$ (4.20)
Sample variance	Sample standard deviation
$s^2 = \frac{\sum_{i=1}^{k} f_i(m_i-\bar{x})^2}{n-1}$ (4.21)	$s = \sqrt{\frac{\sum_{i=1}^{k} f_i(m_i-\bar{x})^2}{n-1}}$ (4.22)

where f_i = frequency or number of observations in the ith group
m_i = midpoint of the ith group
k = number of groups

The shortcut formulas found in Appendix 3 can be used to arrive at the same answer.

Example 4.16 Analyzing a GDP Forecast. Suppose we want to calculate the mean, variance, and standard deviation for a sample of forecasts for next year's GDP growth rate. We record our data in Table 4.10.

Using the shortcut formulas of Eq. 4.21a, we can calculate the sample standard deviation as follows:

$$\bar{x} = \frac{-4 + 12 + 24 + 30}{30} = 62/30 = 2.07$$

$$s^2 = \frac{238 - [30 \times (2.07)^2]}{30 - 1} = \frac{109.45}{29} = 3.77$$

$$s = \sqrt{3.77} = 1.94$$

4.6.5 Percentiles

The calculation of a particular percentile boundary (B) for grouped data for percentiles is similar to that of the median for grouped data:

$$B = L + \frac{(pN - F)}{f}(U - L) \tag{4.23}$$

Here, N is the number of observations, p is the percentile desired, and the product pN gives the corresponding observation. F is the number of observations up to the

Table 4.10 Forecasts of GDP growth rate

Forecast class (%)	Class midpoint (m_i)	Frequency (f_i)	$f_i m_i$	m_i^2	$f_i m_i^2$
−2–0	−1	4	−4	1	4
0–2	1	12	12	1	12
2–4	3	8	24	9	72
4–6	5	6	30	25	150
		30	62		238

Table 4.11 Prime rate, 1990–2009

Class (%)	Frequency	Cumulative frequency
3.1–6	6	6
6.1–9	12	18
9.1–12	2	20
12.1–15	0	20

Source: Economic Report of the President, January 2010

Table 4.12 ACT scores

ACT	x	f	xf	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$	$(x-\mu)^3$	$f(x-\mu)^3$
14–18	16	8	128	−8.5	72.25	578.0	−614.13	−4913.00
19–23	21	34	714	−3.5	12.25	416.5	−42.88	1457.75
24–28	26	20	520	1.5	2.25	45.0	3.38	67.50
29–33	31	10	310	6.5	42.25	422.5	274.63	2746.25
34–38	36	8	288	11.5	132.25	1058.0	1520.88	12,167.00
		80	1,960			2,520		8,610

lower limit of the class that contains the observation, and f is the frequency of the class. L and U are the lower and upper boundaries, respectively.

Example 4.17 Playing with Percentiles for the Prime Rate. Suppose we have the grouped data given in Table 4.11 for the prime interest rate for the past 20 years (1990–2009) [see Table 4.16 in this Chapter].

To determine the 60th percentile boundary, we reason as follows: $p = .60$ and $.60 \times 20 = 12$. The 12th observation is in the 6.1 to 9 class. Thus, $L = 6.10$, $f = 12$, $F = 6$, and $U = 9.00$. Our estimate of the 60th percentile boundary is therefore

$$6.10 + \frac{(.60)(20) - 6}{12}(9.00 - 6.10) = 7.55$$

By this estimate, 60 % of the observations are below 7.55 %.

Example 4.18 Examining the Skewness of ACT Scores. Suppose the grouped data in Table 4.12 represent the ACT scores for a high school class. The population skewness for *grouped* data is

$$\mu_3 = \frac{\sum_{i=1}^{k} f_i(x_i - \mu)^3}{N} \tag{4.24}$$

Using the information listed in Table 4.12, we can calculate the summary statistics as follows:

$$\mu = \frac{1960}{80} = 24.5$$

$$\sigma^2 = \frac{2520}{80} = 31.5$$

$$\sigma = 5.61$$

$$\text{Skewness} = \mu_3 = \sum \frac{f(x-\mu)^3}{N} = \frac{8610}{80} = 107.63$$

$$\text{CS} = \frac{\mu_3}{\sigma^3} = \frac{107.63}{176.56} = .610$$

The skewness μ_3 is positive and equal to 107.63. The coefficient of skewness (.610) is positive, indicating that the distribution is skewed to the right. We can use the same formula (Eq. 4.24) for a sample if we replace μ by $\bar{x}$ and N by n.

4.7 Applications

In this section, we will demonstrate how measures of central tendency, dispersion, position, and skewness can be used to analyze sample market survey data, rates of return on a stock, and economic data.[7] First, a sample of survey data is used to show how statistical analysis can be applied in making an inventory decision. Second, these same concepts are used to examine the market rates of return for Johnson & Johnson and Merck stock and for the stock market overall. The T-bill and prime interest rates are explored in the third application, and the fourth application involves the macroeconomic variables GNP, personal consumption, and disposable income. Finally, in Appendix 3, we return to the seven accounting ratios for the auto industry presented in Chap. 3 and analyze them in terms of mean, median, MAD, variance, standard deviation, and coefficient of variation, as well as percentiles and skewness.

Application 4.1 Statistical Analysis of a Soda Survey. Suppose Jack Miller, a manager at A&P, wants to determine the average monthly purchase, per dwelling unit, of six-packs of soda in the central New Jersey area. This average monthly purchase information will help A&P establish an inventory policy.

[7] Financial ratio analysis for two pharmaceutical firms is carried out in Appendix 3.

Table 4.13 Monthly purchase of six-packs of soda per dwelling unit in the central New Jersey area as of January 1991

i	x_i	i	x_i
1	8	21	9
2	4	22	8
3	4	23	1
4	9	24	4
5	3	25	6
6	3	26	5
7	1	27	4
8	2	28	2
9	0	29	1
10	4	30	0
11	2	31	8
12	3	32	7
13	5	33	5
14	7	34	6
15	10	35	4
16	6	36	3
17	5	37	2
18	7	38	1
19	3	39	0
20	2	40	8

Table 4.14 Summary measures of monthly purchase of six-packs of soda

Arithmetic mean	4.30
Median	4.00
Mode	4.00
Range	10.00
Standard deviation	2.77
Variance	7.65
Mean absolute deviation from mean	2.30
Mean absolute deviation from median	2.25
Coefficient of variation	.64
Coefficient of skewness	1.32
Percentiles	
10th	1.00
25th	2.00
50th	4.00
75th	6.00
90th	8.00
Interquartiles ranges	4.00

Miller has hired you to conduct a survey and perform statistical analyses in accordance with simple random survey procedures. Let us see how you proceed. First you conduct a survey and assemble the results in Table 4.13. Then, using the random sample data, in Table 4.13, you calculate the related summary statistics and list them in Table 4.14. You have computed all the descriptive statistics discussed in this chapter except the geometric mean. (Because some of the values of x are equal

Table 4.15 Rates of return for JNJ, MRK, and S&P 500

Year	JNJ	MRK	S&P 500
1990	0.230	0.185	0.036
1991	0.617	0.879	0.124
1992	−0.551	−0.734	0.105
1993	−0.092	−0.183	0.086
1994	0.245	0.142	0.020
1995	0.585	0.754	0.177
1996	−0.410	0.235	0.238
1997	0.341	0.353	0.303
1998	0.288	0.410	0.243
1999	0.124	−0.537	0.223
2000	0.140	0.412	0.075
2001	−0.431	−0.357	−0.163
2002	−0.078	−0.013	−0.168
2003	−0.021	−0.158	−0.029
2004	0.249	−0.272	0.171
2005	−0.032	0.037	0.068
2006	0.122	0.418	0.086
2007	0.035	0.367	0.127
2008	−0.076	−0.451	−0.174
2009	0.108	0.254	−0.223

to zero, it would make no sense to compute the geometric average, wherein the data are multiplied together.)

Your measures of central tendency—the mean, median, and mode—indicate where the center of the data is. In addition, because the mean is greater than the median, you know the data are positively skewed. The fact that the coefficient of skewness is positive confirms this.

The variance, standard deviation, and mean absolute deviation provide information on how the data are spread out around their average value.

Finally, you show percentiles for the data. The *n*th percentile reveals that *n* percent of the data will be below that value. For example, the 50th percentile has a value of 4, so 50 % of the data will have a value of 4 or less. Likewise, because the 90th percentile has a value of 8, 90 % of the data will have a value of 8 or less. The interquartile range is just the difference between the third and first quartiles.

Using the information you have provided, the manager can make better decisions about how many six-packs of soda to keep in inventory.

Application 4.2 Stock Rates of Return for Johnson & Johnson, Merck, and the Market. Central tendency and dispersion statistics can also be used to analyze the rates of return for JNJ and MRK stock, as well as for the general stock market.

The rates of return listed in Table 4.15 are calculated on a yearly basis. With these statistics, we can determine whether the stock rates of return for the two firms fluctuated more than the market. And we can determine whether the stocks have generally outperformed or underperformed the market over this period.

Much useful information can be obtained by merely perusing Table 4.15. The greatest gain for JNJ occurred in 1991, when the price of its stock increased by .617 %. JNJ's worst year occurred in 1992, when its stock lost .551 % of its value.

The range of the three stock rates of return—the difference between the highest and lowest values—was 1.168 (.617 + .551) for JNJ, 1.62 (.879 + .734) for MRK, and .526 (.303 + .223) for the overall market. Note that the three tended to move together; when the market went up, the stocks also tended to rise and vice versa. However, this does not always hold, as can be seen in 2009 when the market went down by over .233 % and the two pharmaceutical stocks went up. The relationship between the market rate of return and the rate of return for individual firms will be analyzed in Chap. 14 when simple regression analysis is discussed.

Examining the ranges alone makes it appear that the overall market was less volatile than the two stocks. However, recall that because the range takes into consideration only the highest and lowest observations, it is strongly influenced by outlying observations. Therefore, we must examine more sophisticated measures of dispersion, such as the standard deviation, CS, and CV, to obtain a sense of the volatility of the observations. MRK had the highest standard deviation (.425), followed by JNJ (.303) and the market (.151). These rankings hold when comparing the coefficient of variation (CV) as well. Here, MRK has the highest CV followed by JNJ and the market.

For the two pharmaceutical firms, the mean return is less than the median due to the fact that a few extreme observations on the low side of the distribution are pushing down the mean. For example, JNJ had a return of 12.4 % in 1999, while MRK suffered a return of −53.7 % in 1999. These observations affect the mean but do not influence the median. Thus, the median is probably a more accurate indicator of central tendency.

Investors would have preferred owning MRK instead of JNJ stock because of its mean and median returns of 8.7 % and 16.4 %, respectively. However, when MRK is compared to the market, notice that MRK has a higher mean but a lower median than the market. We also notice that both JNJ's and Merck's means are lower than their medians, indicating their returns are negatively skewed, matching the market's mean below its median, indicating negative skewness. This indicates that MRK, JNJ, and the market all have more observations which lie above the mean. See Sect. 9.7 in Chap. 9 for further discussion on this implication.

MEAN	0.070	0.087	0.066
MEDIAN	0.115	0.164	0.086
STD	0.303	0.425	0.151
CS.	-10.688	-1.743	-162.890
CV	4.353	4.879	2.284
PERCENTILES			
10th	-0.412	-0.459	-0.168
25th	-0.077	-0.205	0.008
50th	0.115	0.164	0.086
75th	0.246	0.378	0.173
90th	0.365	0.452	0.238

Table 4.16 T-bill rate and prime rate (%), 1990–2009

Year	3-Month T-bill rate	Prime Rate
90	7.49	10.01
91	5.38	8.46
92	3.43	6.25
93	3.00	6.00
94	4.25	7.14
95	5.49	8.83
96	5.01	8.27
97	5.06	8.44
98	4.78	8.35
99	4.64	7.99
00	5.82	9.23
01	3.39	6.92
02	1.60	4.68
03	1.01	4.12
04	1.37	4.34
05	3.15	6.19
06	4.73	7.96
07	4.35	8.05
08	1.37	5.09
09	0.15	3.25
Mean	3.7735	6.9785
Median	4.3	7.55
Std dev	1.8952	1.9018
CV	0.4407	0.2725
CS	−0.0398	−0.0671

Application 4.3 3-Month Treasury Bill Rate and Prime Rate. Table 4.16 shows two key interest rates, the 3-month T-bill rate and the prime rate, for the period 1990–2009. As we have noted, a T-bill is a short-term debt instrument issued by the United States government. T-bills are backed by the full faith and credit of the US government, which makes these investments the safest in the world and the closest thing to a risk-free asset. The prime rate is the rate that banks charge their best customers, such as large corporations. This rate may differ slightly from bank to bank.

As Table 4.16 shows, the T-bill rate fluctuated between 3 % and 7 % from 1990 to 2001. It stayed constant with a little decline in 2002 until 2004, with a slight increase again to 3.15 % in 2005. The trend for the prime rate is similar, although the prime is several percentage points higher.

The T-bill rate is lower largely because the T-bill is close to a risk-free asset, whereas the prime rate includes a risk premium. This relationship is illustrated statistically by the means and medians of the two rates. The mean for the T-bill rate from 1990 to 2009, for example, is 3.77 %; it is 6.97 % for the prime rate. For the same period, the median for the T-bill rate is 4.3 %; it is 7.55 % for the prime rate. The fact that the mean is lower than the median indicates that a few extreme observations are pushing down the mean and that both distributions are negatively

skewed. The coefficient of skewness for the T-bill rate is −3.98 %, which is higher than that for the prime rate, −6.7 %.

The range for the T-bill over the period 1990–2009 is 7.49 − 0.15 = 7.34; the range for the prime is 10.01 − 3.25 = 6.76. The standard deviation for the prime (1.9018) is also higher than the T-bill standard deviation (1.8952). Finally, the coefficient of variation is higher for the prime rate. One reason why the prime rate has fluctuated more than the 3-month T-bill rate during this 20-year period may be the fact that 3-month treasury bills, in general, have shorter maturities than loans made at the prime rate. Because this time period was marked by more volatile interest rates (especially between 2000 and 2006), the prime rate may have been adjusted to reflect the added risk associated with longer-term loans. In addition, because T-bills are marketable securities, the rate they return is determined by the market. Loans made at the prime rate, in contrast, arc not marketable, so the prime rate is adjusted (by bankers) only periodically. This may make it more volatile.

Application 4.4 GDP, Personal Consumption, and Disposable Income. Here, we will examine annual data on national income. The GDP is one of the most popular economic indicators because it measures the market value of all final goods and services produced in the United States within a given time period. GDP is a key indicator of the health of the economy: the occurrence of consecutive quarters of decline in real (inflation-adjusted) GDP is sometimes used to define a recession. The GDP is calculated by adding personal consumption expenditures, gross private domestic investment, government purchases, and net exports. Disposable income is the amount of after-tax income that individuals have available to spend.

Data on GDP, personal consumption, and disposable income in constant 2005 dollars for the period 1960–2009 are shown in Table 4.17. Clearly, there is a tendency for the indicators to increase steadily disposable income declined in only 1 year (1972–1973), while personal consumption only declined in 2009. GDP declined in 6 years (1974, 1975, 1980, 1982, 1991, and 2009. The mean for GDP is $7339.92 billion dollars, and the median is $6713.2. Because the mean is greater than the median, the distribution is positively skewed. The mean is also greater than the median for consumption and disposable income. However, there is a large difference between the two measures of central tendency for GDP, personal consumption, and disposable income, indicating that there is a great deal of skewness and, therefore, the distribution is not symmetric.

The standard deviation for GDP is higher than the standard deviation for consumption. However, the CV is lower for GDP. The CV is probably a better measure of variability than the standard deviation because the level of GDP is always higher than consumption. The standard deviation for disposable income is lower than the standard deviation for GDP and greater than the standard deviation for consumption. However, the CV of disposable income is lower than the CV for consumption and greater than

Table 4.17 GDP, personal consumption, and disposable income, 1960–2009

Year	GDP	Personal consumption	Disposable income
1960	2,830.90	331.8	365.2
1961	2,896.90	342.2	381.6
1962	3,072.40	363.3	404.9
1963	3,206.70	382.7	425
1964	3,392.30	411.5	462.3
1965	3,610.10	443.8	497.8
1966	3,845.30	480.9	537.4
1967	3,942.50	507.8	575.1
1968	4,133.40	558	624.7
1969	4,261.80	605.1	673.8
1970	4,269.90	648.3	735.5
1971	4,413.30	701.6	901.4
1972	4,647.70	770.2	869
1973	4,917.00	852	978.1
1974	4,889.90	932.9	1,071.7
1975	4,879.50	1,033.80	1,187.3
1976	5,141.30	1,151.30	1,302.3
1977	5,377.70	1,277.80	1,435
1978	5,677.60	1,427.60	1,607.3
1979	5,855.00	1,591.20	1,790.9
1980	5,839.00	1,755.80	2,002.7
1981	5,987.20	1,939.50	2,237.1
1982	5,870.90	2,075.50	2,412.7
1983	6,136.20	2,288.60	2,599.8
1984	6,577.10	2,501.10	2,891.5
1985	6,849.30	2,717.60	3,079.3
1986	7,086.50	2,896.70	3,025.8
1987	7,313.30	3,097.00	3,435.3
1988	7,613.90	3,350.10	3,726.3
1989	7,885.90	3,594.50	3,991.4
1990	8,033.90	3,835.50	4,254
1991	8,015.10	3,980.10	4,444.9
1992	8,287.10	4,236.90	4,736.7
1993	8,523.40	4,483.60	4,921.6
1994	8,870.70	4,750.80	5,184.3
1995	9,093.70	4,987.30	5,457
1996	9,433.90	5,273.60	5,759.6
1997	9,854.30	5,570.60	6,074.6
1998	10,283.50	5,918.50	6,498.9
1999	10,779.80	6,342.80	6,803.3
2000	11,226.00	6,830.40	7,327.2
2001	11,347.20	7,148.80	7,684.5
2002	11,553.00	7,439.20	8,009.7
2003	11,840.70	7,804.00	8377.8
2004	12,263.80	8,285.10	8,889.4
2005	12,638.40	8,819.00	9,277.3

(continued)

Table 4.17 (continued)

Year	GDP	Personal consumption	Disposable income
2006	12,976.20	9,322.70	9,915.7
2007	13,254.10	9,826.40	10,403.1
2008	13,312.20	10,129.90	10,806.4
2009	12,988.70	10,092.60	10,964.5
Median	6,713.2	2,609.35	2,958.65
Mean	7,339.924	3,522.16	3,840.374
Std Dev	3,225.559	3,077.678	3,276.337
CS	0.417176	0.744179	0.708301
CV	0.439454	0.873804	0.85313

the CV for GDP. Thus, we must use the CV when comparing the variability of data that are different in range of values. Because GDP is greater than consumption and disposable income, it usually has a greater variance. But using the CV makes it possible to compare the dispersion because the dispersion is standardized.

4.8 Summary

In this chapter, we showed how a series of data can be described by using only a few summary statistics.

1. Measures of central tendency such as the mean, median, and mode provide information on the center of the distribution.
2. Measures of dispersion such as the variance, standard deviation, mean absolute deviation, and coefficient of variation provide information on how spread out the data are.
3. Measures such as percentiles, quartiles, interquartiles, and Z scores provide information on the relative position of a data set.
4. Shape measures such as skewness are used to measure the degree of a distribution's asymmetry.

In the next chapter, we introduce the concepts of probability that are required for making statistical inference. However, we will continue to use descriptive statistics such as the mean and variance to describe the central tendency and the dispersion of a distribution.

Questions and Problems

1. The midterm scores from an honors seminar in accounting are 25, 84, 82, 83, 90, 91, 99, 100, and 100. Find the mean, median, and mode. Is one measure preferable to another? Why or why not?
2. What is your mean speed if you drive 35 miles per hour for 2 h and 55 miles per hour for 3 h?

3. What are descriptive statistics? Why are they important? Give some examples of descriptive statistics.
4. The following sample annual starting salaries were offered to 12 college seniors in 1992:

$21,400	$15,600	$16,500	$24,200
22,300	20,000	17,000	21,750
18,750	19,250	14,900	15,750

   (a) Calculate the mean and median for these observations.
   (b) Calculate the variance and standard deviation for these observations.
   (c) Use MINITAB to construct a box and whisker plot and explain the result.

5. A $250 suit is on sale for $190, and a $90 pair of shoes is on sale for $65. Find the average percent decrease in price for the 2 items.
6. The following are the average daily reported share volumes traded on the NYSE, in thousands, for the years listed:

1971	15,381	1981	46,882
1972	16,487	1982	64,859
1973	16,084	1983	85,336
1974	13,904	1984	91,229
1975	18,551	1985	109,132
1976	21,186	1986	141,489
1977	20,928	1987	188,796
1978	28,591	1988	161,509
1979	32,233	1989	165,568
1980	44,867	1990	156,777

   (a) Calculate the mean and median share volume for these observations.
   (b) Calculate the variance and standard deviation for these observations.

7. The following data are annual rates of return on the DJIA and the S&P 500:

	DJIA	S&P 500
1960	−9.34	−2.97
1961	18.71	23.13
1962	−10.91	−11.81
1963	17.12	18.89
1964	14.57	12.97
1965	10.88	9.06
1966	−18.94	−13.09
1967	15.20	20.09
1968	5.24	7.66
1969	−15.19	−11.36
1970	4.82	.10
1971	6.11	10.80

(continued)

	DJIA	S&P 500
1972	14.58	15.57
1973	−16.58	−17.37
1974	−27.57	−29.64
1975	38.34	31.49
1976	17.86	19.18
1977	17.27	−11.53
1978	−3.15	1.05
1979	4.19	12.28
1980	5.57	25.86
1981	4.66	−9.94
1982	−5.21	15.49
1983	34.60	17.06
1984	−1.00	1.15
1985	12.71	26.33
1986	34.97	14.62
1987	26.95	2.03
1988	−9.45	12.40
1989	21.74	27.25
1990	6.78	−6.56

(a) Calculate the arithmetic mean and standard deviation of the DJIA and the S&P 500 for the years 1960–1979, 1970–1989, and 1981–1990.
(b) Calculate the geometric mean for these same years.

8. A sample of 20 workers in a small company earned the following weekly wages:
$175, 175, 182, 175, 175, 200, 250, 225, 250, 200, 195, 200, 200, 190, 325, 300, 310, 325, 400, 225

(a) Calculate the mean and standard deviation.
(b) Calculate the mode.
(c) Calculate the median.

9. You are given the following information about stock A and stock B:

		Returns next year	
State of world next year	Chance of occurrence	A	B
Recession	.30	10 %	9.8 %
Normal growth	.40	11 %	11.2 %
Inflation	.30	12 %	13.0 %

(a) Calculate the mean, standard deviation, and coefficient of variation for each stock.
(b) If you could purchase only one stock, which would you choose? Why?

10. Consider the following annual data on profit rates for Cherry Computers, Lemon Motors, and Orange Electronics:

Year	Cherry computers	Lemon motors	Orange electronics
1983	14.2	−6.2	37.5
1984	12.3	13.3	−10.6
1985	−16.2	−8.4	40.3
1986	15.4	27.3	5.4
1987	17.2	28.2	6.2
1988	10.3	14.5	10.2
1989	−6.3	−2.4	13.8
1990	−7.8	−3.1	11.5
1991	3.4	15.6	−6.2
1992	12.2	18.2	27.5

(a) Calculate the mean and standard deviation of each company's profits.
(b) Compare the performance of these three companies. Which company do you believe was the best performer over these 10 years?

11. The following table gives the price of Charleston Corporation's stock under different economic conditions:

Economic condition	Chance of occurrence	Price per share
Depression	.25	$65
Recession	.25	$80
Normal growth	.3	$95
Inflation	.2	$100

(a) Sketch a relative frequency diagram for Charleston's stock.
(b) Calculate the mean and standard deviation of the stock's price.

12. The final scores from an honors seminar in marketing were
65 55 70 80 90 100 50 75
Find the mean, median, and mode. Is one measure preferable to another. Why or why not?

13. (a) Briefly compare the arithmetic mean with the geometric mean. Cite some cases where the geometric mean would be preferred.
(b) Use data given in Table 4.9 to calculate the arithmetic mean and the geometric mean of market rates of return during 1970–1990.
(c) Analyze the results which you obtained in part (b).

14. Compare the use of the mean to the use of the median as a measure of central tendency. If you were taking a tough calculus class where 3 brilliant students out of 20 nevertheless received perfect scores of 100 on the midterm, would you prefer that the professor use the mean or the median to determine the average grade? Or would it make no difference? (*Hint*: If you said it doesn't matter, you must be very good at calculus; you're one of the three who got a perfect score!)
15. In major league baseball, rookies earn a minimum salary of $100,000, whereas superstar players earn as much as $5 million per year. Do you think the mean or the median of major league salaries would be higher?
16. Suppose you are a market researcher and have been asked to assess the popularity of four brands of coffee. Should you construct your test on the basis of the mean, median, or mode?
17. Why is the standard deviation sometimes preferred to the variance as a measure of dispersion, even though they measure the same thing?
18. In finance, we generally use a measure of dispersion such as the variance to measure the risk of a stock's returns. Explain why the variance may not, however, be the best measure of risk of a stock's returns.
19. Carefully explain the difference between a population and a sample. Why is the formula we use to calculate the population standard deviation different from the one we use to calculate the sample standard deviation?
20. The members of the offensive line of the Denver Broncos weigh 275, 281, 285, 265, and 292 lb, respectively.

    (a) Calculate the mean and median weight of the offensive line.
    (b) Calculate the variance and standard deviation for these observations.

21. A quality control manager finds the following number of defective light bulbs in 10 cases of light bulbs:

Case	Number defective	Case	Number defective
1	3	6	6
2	3	7	3
3	7	8	4
4	1	9	5
5	0	10	2

    (a) Draw a frequency diagram for the class intervals
        0–2 3–5 6–8 9 and over
    (b) Draw a relative frequency diagram for these data.
    (c) Calculate the mean, variance, and skewness for the observations and do some analysis.

22. Use the data of question 21 to calculate

    (a) Coefficient of skewness
    (b) Pearson coefficient

23. You are given the following information about two stocks:

Economic condition	Chance of occurrence	Return on A (%)	Return on B (%)
Recession	.25	7	0
Normal growth	.50	8	10
Inflation	.25	9	20

    (a) Calculate the mean, standard deviation, and coefficient of variation for each stock.
    (b) If you had to purchase only one stock, which would you choose?

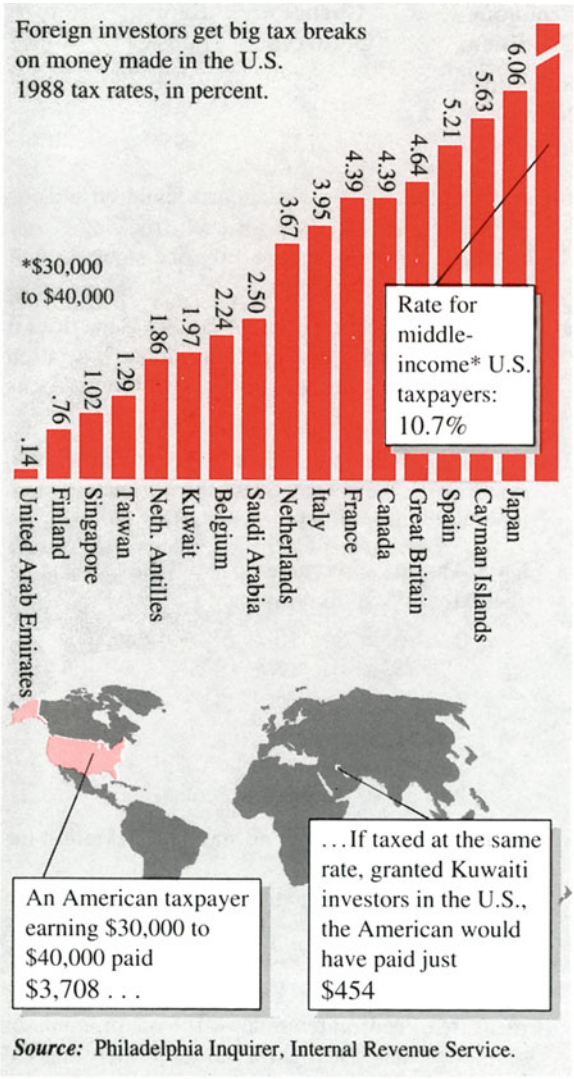
24. What is the coefficient of variation? What does it measure? Explain how the coefficient of variation can be used to decide which of these two stocks to purchase.
25. Suppose you are an efficiency expert who is concerned with the absentee rate for workers in a factory. You collect the following information:

Days absent per month	Number of employees
0	10
1	17
2	25
3	28
4	30
5	27

    (a) Calculate the moan and standard deviation for days absent.
    (b) Calculate the median and mode of the distribution.
    (c) Is the distribution symmetric?

26. When a distribution is skewed to the right, which measure of central tendency—the mean, median, or mode—has the highest value? Which has the lowest value?
27. Calculate the mean, variance, and skewness coefficient for the data given in Table 4.9. Is the distribution symmetric ?
28. On November 17, 1991, the *Home News* used the information in this figure to show that the US Congress taxes foreigners at lower rates than it taxes American citizens.

    (a) Calculate the mean and standard deviation of the tax rates for the 16 foreign countries.
    (b) Calculate the Z scores for the 16 foreign countries and then draw related conclusions.



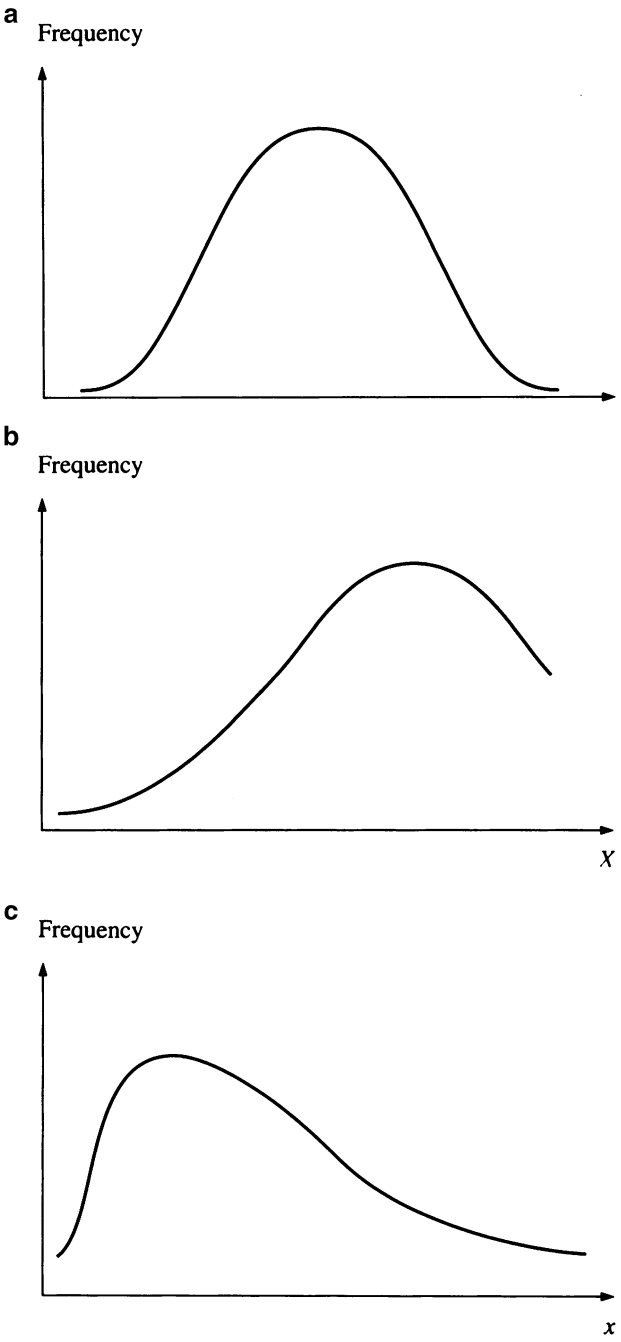
Congress Taxes Foreigners at Lower Rates Than U.S. Citizens

Source: Reprinted by permission of Knight-Ridder Tribune News

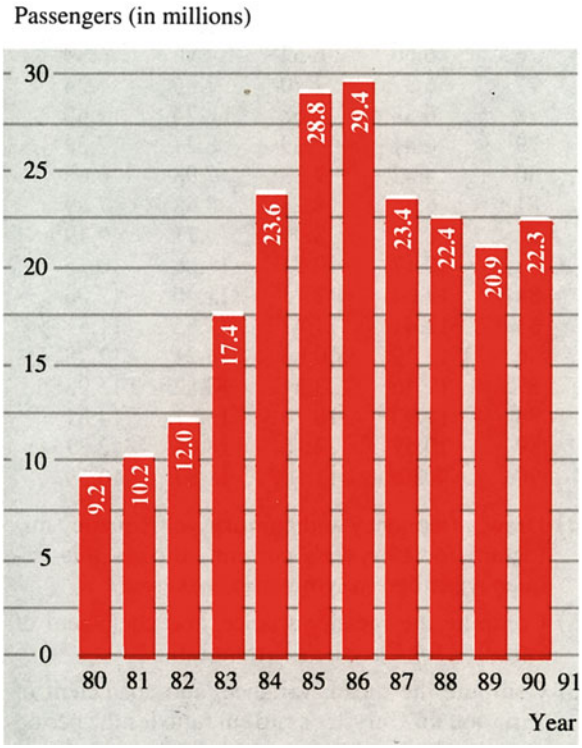
29. Compare the following measures of dispersion: variance, standard deviation, mean absolute deviation, and range.

    (a) What are the benefits and disadvantages of each measure?
    (b) Which measure is the easiest to compute? Which is the most difficult?

30. Calculate the range and mean absolute deviation for the data given in question 21.
31. Explain whether each of the following distributions is symmetric, positively skewed, or negatively skewed.
32. (a) Use the data from the following figure, which are reprinted from the *Home News* of November 17, 1991, to calculate the mean, standard deviation, and Z score of Newark International Airport's passenger traffic trends for the period 1980–1990.
    (b) What do the Z scores you obtained in part (a) suggest?



Source: Port Authority of NY and NJ



33. Your long-lost great uncle has recently died, leaving you $5,000 but stipulating that you must invest it in the stock of either XYZ Company or ABC Company. To compare their rates of return you calculate that over the last 10 years, they had the following means and standard deviations:

	XYZ (%)	ABC (%)
Mean	8	10
Standard deviation	2	3

Which stock would you choose? Have you done everything possible to help you make this decision? That is, are there any other statistics that would be helpful?

34. Suppose you were the agent for Ralph "Boomer" Smith, the punter for the Los Angeles Rams. Explain how you could use the mean yards and standard deviation for Boomer's punts to argue for a pay increase for Boomer.
35. You are a quality control specialist for Brite Lite Company, a light bulb manufacturer. Carefully explain why the standard deviation is as important to you as the mean number of defective light bulbs per case.
36. Use the data given in question 23 of Chap. 2 to compute the mean, standard deviation, and coefficient of variation for

(a) The exchange rate between dollars and pounds.
(b) The exchange rate between dollars and yen.

37. Comment on the following statement: "Investors don't care about the variability of a stock's returns, because they have the same chance of falling below the median as above the median. Therefore, on average, their returns will be the same."
38. (a) Use the 3-month T-bill rate and prime rate information given in Table 4.16 to calculate the Z score.
    (b) Use this information on Z score to do related analysis.
39. Use the data given in Table 3.9 to compute the mean, standard deviation, coefficient of variation, and coefficient of skewness for the current ratio of JNJ.
40. Repeat question 39 using the current ratio data for MRK.
41. You would like to compare the risk and return of two mutual funds. You have the following information:

	Fund A (%)	Fund B (%)
Expected return	10	7
Standard deviation	3	2.5

Which fund do you think is more desirable? Explain.

42. You are given the following information about two stocks:

State of economy	Chance of occurrence	Return on A (%)	Return on B (%)
Poor	.35	5	0
Good	.20	6	10
Excellent	.45	9	20

(a) Calculate the mean and standard deviation for each stock.
(b) Compare the mean, standard deviation, and coefficient of variation of each stock. Is the coefficient of variation or the standard deviation a better measure of risk here?
(c) If you could buy only one stock, which would you choose?

43. Calculate the skewness coefficient for the data given in question 42.

The following information is for questions 44–52. The following table gives the current ratio and inventory turnover for Chrysler, Ford, GM, and the auto industry from 1969 to 1990:

		Current ratio		
Year	Chrys	Ford	GM	Indus
69	1.32	1.37	2.30	1.66
70	1.40	1.33	1.93	1.55
71	1.46	1.39	1.76	1.54

(continued)

(continued)

		Current ratio		
Year	Chrys	Ford	GM	Indus
72	1.49	1.44	2.12	1.68
73	1.55	1.37	2.04	1.65
74	1.36	1.28	1.91	1.52
75	1.27	1.33	1.99	1.53
76	1.37	1.37	1.95	1.56
77	1.34	1.38	1.92	1.55
78	1.43	1.33	1.79	1.52
79	.97	1.25	1.68	1.30
80	.94	1.04	1.26	1.08
81	1.08	1.02	1.09	1.06
82	1.12	.84	1.13	1.03
83	.80	1.05	1.40	1.08
84	.97	1.11	1.36	1.15
85	1.12	1.10	1.09	1.10
86	1.05	1.18	1.17	1.13
87	1.74	1.24	1.53	1.50
88	1.76	1.29	1.71	1.59
89	1.59	.97	1.72	1.43
90	1.50	.94	1.37	1.27

		Inventory turnover		
Year	Chrys	Ford	GM	Indus
69	5.76	6.46	6.46	6.27
70	5.03	6.03	4.56	5.21
71	5.68	6.47	7.08	6.41
72	7.11	7.26	7.25	7.21
73	6.53	6.41	6.92	6.62
74	4.47	5.55	4.93	4.98
75	5.61	6.31	6.28	6.07
76	6.60	6.62	7.46	6.89
77	6.37	7.70	7.66	7.24
78	6.88	7.58	8.34	7.60
79	6.41	7.39	8.21	7.34
80	4.82	7.23	7.98	6.68
81	6.76	8.24	8.68	7.89
82	8.87	8.99	9.71	9.19
83	10.17	10.81	11.26	10.75
84	12.04	12.73	11.40	12.06
85	11.41	11.47	11.65	11.51
86	13.29	10.83	14.21	12.78
87	11.46	23.51	12.82	15.93
88	11.93	18.70	13.81	14.81
89	10.57	12.16	14.08	12.27
90	8.39	11.50	11.87	10.59

44. Draw a frequency and cumulative frequency histogram for Chrysler's current ratio. Is this frequency histogram symmetric or skewed?
45. Compute the mean, variance, and coefficient of variation for Chrysler's current ratio.
46. Compute the mean, variance, and coefficient of variation for Chrysler's current ratio for the period 1969–1978 and the period 1979–1990. Compare these descriptive statistics to those you computed in question 45. Have any changes occurred over the two different time periods? If so, can you propose an explanation?
47. Draw a frequency and cumulative frequency histogram for Chrysler's inventory turnover. Is this frequency histogram symmetric or skewed?
48. Compute the mean, variance, and coefficient of variation for Chrysler's inventory turnover.
49. Compute the mean, variance, and coefficient of variation for Chrysler's inventory turnover for the period 1969–1978 and the period 1979–1990. Compare these descriptive statistics to those you computed in question 48. Have any changes occurred over the two different time periods? If so, can you propose an explanation?
50. Answer the following questions by referring to the MINITAB output of Ford's current ratio:

    (a) Is the frequency histogram of Ford's current ratio symmetric or skewed?
    (b) Compare and analyze the means, variances, and coefficients of variance of Ford's current ratio calculated from different periods.

51. Using a calculator and the MINITAB program, answer questions 45–46 and 48–49 again, using the data for GM.
52. Answer questions 45–46 and 48–49 again, using the data for the auto industry.

The following information from *Best's Aggregates and Averages* can be used for questions 53–55. You are given the following information on the property–casualty insurance industry. NPW (net premiums written) is a measure of the dollar value of premiums written in property–casualty insurance (such as auto insurance, home insurance, and so on). PHS (policyholders' surplus) is a measure of the net worth, or equity, of an insurer:

Insurance industry		
Year	NPW	PHS
1967	23,583	14,802
1968	25,766	16,192
1969	28,956	13,964
1970	32,578	15,499

MINITAB Output of Ford's Current Ratio (Question 50)

```
MT3 > NAME C1 'FORD'
MTB > SET INTO 'FORD'
```

```
DATA> 1.379 1.333 1.249 1.044 1.024 0.844 1.049 1.114
  1.097 1.181
DATA> 1.235 1.488 1.480 1.453 1.539 1.552 1.586 1.701
  1.674 1.688
DATA> END
MTB > GSTD
* NOTE * Standard Graphics are enabled.
  Professional Graphics are disabled.
Use the GPRO command to enable Professional Graphics.
MTB > HISTOGRAM 'FORD'
```

Character Histogram

Histogram of FORD N = 20

Midpoint	Count	
0.8	1	*
0.9	0	
1.0	3	* * *
1.1	2	* *
1.2	3	* * *
1.3	1	*
1.4	1	*
1.5	4	* * * *
1.6	2	* *
1.7	3	* * *

```
MTB > SET INTO 'FORD'
DATA. 1.379 1.333 1.249 1.044 1.024 0.844 1.049 1.114
  1.097 1.181
DATA> END
MTB > MEAN 'FORD'
```

Column Mean

Mean of FORD = 1.1314

MTB > STDEV 'FORD'

Column Standard Deviation

Standard deviation of FORD = 0.15926

```
MTB > NAME C2 'F0RD2'
MTB > SET INTO 'FORD2'
DATA> 1.235 1.488 1-480 1.453 1.539 1.552 1.586 1.701
  1.674 1.688
DATA> END
MTB > MEAN 'FORD2'
Column Mean
Mean of FORD2 = 1.5396
MTB > STDEV 'FORD2'
```

Column Standard Deviation

Standard deviation of FORD2 = 0.13942

Year	NPW	PHS
1971	35,860	19,065
1972	38,930	23,812
1973	42,075	21,389
1974	44,704	16,270
1975	49,605	19,712
1976	60,439	24,631
1977	72,406	29,300
1978	81,699	35,379
1979	90,169	42,395
1980	95,702	52,174
1981	99,373	53,805
1982	104,038	60,395
1983	109,247	65,606
1984	118,591	63,809
1985	144,860	75,511
1986	176,993	94,288

53. Draw a relative and cumulative relative frequency histogram for NPW and PHS.
54. Compute the mean, standard deviation, and coefficient of variation for NPW and PHS.
55. If you were interested in comparing the dispersion of NPW to the dispersion of PHS, should you use the variance, the standard deviation, or the coefficient of variation'? Which variable—NPW or PHS—has the greater dispersion around its mean?
56. Suppose the variance of a population is 0. What can you say about the members of that population?
57. Suppose you have three populations containing two members each. Suppose the means and the variances of the three populations are the same. Are the numerical values of the members of the first population necessarily identical to the numerical values of the members of the second or third population?
58. Reconsider question 57, but this time assume that each population has three members.
59. In a class of ten students, we find that the students spent the following amounts of money on textbooks for the semester:

$225	$178	$272	$310	$190	$145	$150
$220	$285	$112				

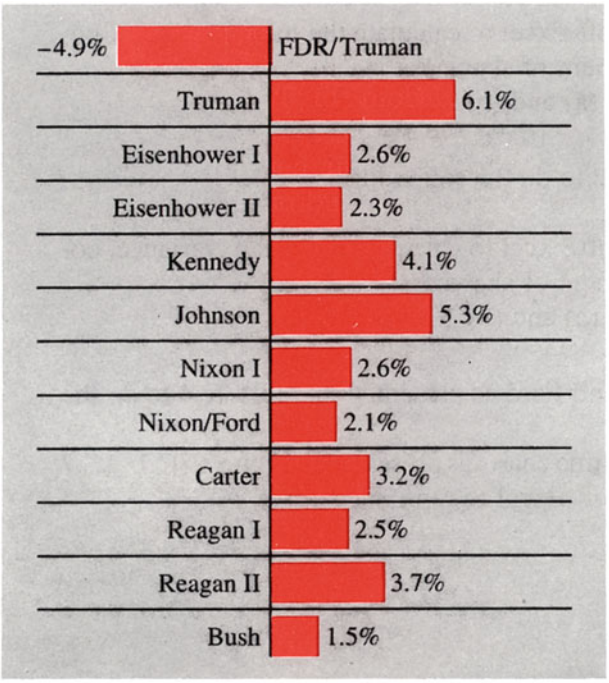
(a) Find the median dollar value spent on books.
(b) Find the mean and standard deviation for the dollar value spent on books.

60. Suppose you are in a statistics class of 12 students. Your score on the test is a 75 out of 100. The scores for the entire class (including your score) were

100	75	100	50	45	100	60	65	72	70	66	74

Would you prefer that the teacher use the median or the mean as the average score when deciding how to draw the curve that she or he will use in determining grades?

61. The *Home News* used the chart reproduced here in its September 29, 1991, issue to show the economic growth record for 12 different periods since World War II.



Annual Average GNP Growth by Presidential Term Since World War II

Source: Knight-Ridder Tribune News/Marty Westman and Judy Treible as found in the *Home News.* September 29, 1991. Reprinted by permission of Knight-Ridder Tribune News

(a) Calculate the mean and the median.
(b) Calculate the standard deviation and mean absolute deviation.
(c) Calculate the Z scores and do related analysis.

62. The following table, reprinted from the November 20, 1991, *Wall Street Journal*, shows the percentage change of stock prices for ten Dow Jones sectors over two different periods.

    (a) Calculate the arithmetic mean and geometric mean of the two sets of data.
    (b) Calculate the standard deviation and mean absolute deviation of these two sets of data.
    (c) Calculate the Z scores and do related analysis.

Of the market's slide: performance of the DJ sectors

	% Change 11/13/91 to 11/19/91	% Change 12/31/90 to 11/19/91
Conglomerates	−2.73 %	19.50 %
Utilities	−3.50	3.73
Energy	−3.65	2.63
Consumer noncyclical	−3.97	24.35
DJ equity index	−4.52	16.53
Technology	−5.04	15.32
Consumer cyclicals	−5.16	23.11
Industrial	−5.36	13.94
Basic materials	−5.63	15.64
Financial	−5.73	31.12

Source: Wall Street Journal, November 20, 1991. Reprinted by permission of the Wall Street Journal, © 1991 Dow Jones & Company, Inc. All Rights Reserved Worldwide

Refer to Table 4.17, in which the GDP, personal consumption, and disposable income during the period 1960–2009 are given.

63. Please give the two quartiles Q1 and Q3 for the GDP, personal consumption, and disposable income, respectively.
64. Please draw the box and whisker plots for the GDP, personal consumption, and disposable income, respectively.
65. Please calculate the skewness for the GDP, personal consumption, and disposable income, respectively.
66. Please calculate the kurtosis for the GDP, personal consumption, and disposable income, respectively.

Refer to Table 2.3, in which data on EPS, DPS, and PPS for JNJ, Merck, and S&P 500 during the period 1988–2009 are given.

67. Please calculate the mean, median for EPS, DPS, and PPS for JNJ during the period 1988–2009.
68. Please calculate the standard deviation for EPS, DPS, and PPS for JNJ during the period 1988–2009.

Refer to Table 2.4, in which data on rates of return for JNJ, Merck, and S&P 500 during the period 1989–2009 are given.

69. Please calculate the skewness for the rates of return for JNJ and S&P 500 during the period 1989–2009.
70. Please calculate the kurtosis for the rates of return for JNJ and S&P 500 during the period 1989–2009.

Project I: Project for Descriptive Statistics

1. Use 3-month T-bill rate and prime rate presented in Table 2.1 to do the following:
   (a) Draw a time chart.
   (b) Use either MINITAB or Microsoft Excel to calculate the mean, variance, coefficient of variation, and coefficient of skewness
   (c) Analyze the statistical results of (a) and (b)
2. Use the data presented in Table 2.4 to do the following:
   (a) Draw a time chart.
   (b) Use either MINITAB or Microsoft Excel to calculate the mean, variance, coefficient of variation and coefficient of skewness
   (c) Analyze the statistical results of (a) and (b)
3. Use seven key ratios for both JNJ and Merck as presented in Table 2.8 to do the following:
   (a) Use Microsoft Excel to draw the time charts as presented in Figs. 2.17, 2.18, 2.19, 2.20, 2.21, 2.22, and 2.23
   (b) Use either MINITAB or Microsoft Excel to reproduce Table 4.18
4. Update seven key ratios for both JNJ and Merck as presented in Table 2.8 from Yahoo Finance and redo 3(a) and 3(b).

Appendix 1: Shortcut Formulas for Calculating Variance and Standard Deviation

Population variance	Population standard deviation
$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$ (4.5a)	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2}$ (4.6a)
Sample variance	Sample standard deviation
$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$ (4.7a)	$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}}$ (4.8a)

where N = number of observations in the population
n = number of observations in the sample
$\bar{x}$ = sample mean of x
μ = population mean of x

Appendix 2: Shortcut Formulas for Calculating Group Variance and Standard Deviation

Population variance	Population standard deviation
$\sigma^2 = \frac{\sum_{i=1}^{k} f_i m_i^2}{N} - \mu^2$ (4.19a)	$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i m_i^2}{N} - \mu^2}$ (4.20a)
Sample variance	Sample standard deviation
$s^2 = \frac{\sum_{i=1}^{k} f_i m_i^2}{n-1} - \frac{n\bar{x}^2}{n-1}$ (4.21a)	$s = \sqrt{\frac{\sum_{i=1}^{k} f_i m_i^2}{n-1} - \frac{n\bar{x}^2}{n-1}}$ (4.22a)

where f_i = frequency or number of observations in the ith group
m_i = midpoint of the ith group
k = number of groups

Appendix 3: Financial Ratio Analysis for Two Pharmaceutical Firms

Summary statistics for seven accounting ratios for the two US pharmaceutical companies over the time period 1990–2009 are presented in Table 4.18. The mean and median enable us to compare the central tendencies of the various ratios. Recall that the mean is calculated by adding all the observations in the sample and dividing by the number of observations in the sample (here, 20). That is

Table 4.18 Statistics for selected financial ratios of two pharmaceutical companies

	CA/CL	Inventory turnover	TD/TA	NI/SAL	NI/TA	P/E ratio	Payout ratio
JNJ							
Mean	1.781	2.660	0.448	0.155	0.142	23.208	0.385
Median	1.775	2.496	0.448	0.159	0.145	22.512	0.367
STD	0.301	0.301	0.056	0.037	0.020	6.663	0.053
Variance	0.091	0.091	0.003	0.001	0.000	44.392	0.003
Skewness	0.525	0.360	0.167	−0.305	−0.883	0.415	2.552
Kurtosis	0.823	−1.724	−0.076	−0.570	2.065	−0.694	7.897
MAD	0.215	0.276	0.042	0.031	0.142	5.397	0.034
CV	0.169	0.113	0.125	0.240	0.141	0.287	0.136
CS	19.247	13.180	957.997	−5983.078	−111526.6	0.001	17590.079
Percentiles							
10th	1.495	2.350	0.388	0.116	0.120	16.188	0.344
25th	1.614	2.426	0.422	0.128	0.130	17.556	0.358
50th	1.775	2.496	0.448	0.159	0.145	22.512	0.367
75th	1.867	2.992	0.483	0.181	0.153	27.616	0.387
90th	2.177	3.080	0.535	0.203	0.160	31.681	0.435
	CA/CL	Inventory turnover	TD/TA	NI/SAL	NI/TA	P/E ratio	Payout ratio
Merck							
Mean	1.355	3.589	0.455	0.220	0.244	21.591	0.151
Median	1.309	2.229	0.473	0.200	0.250	20.541	0.162
STD	0.212	2.561	0.066	0.076	0.057	8.409	0.039
Variance	0.045	6.559	0.004	0.006	0.003	70.711	0.002
Skewness	0.438	0.953	−0.085	2.105	−0.279	0.115	−0.152
Kurtosis	−0.364	−0.261	−1.200	5.672	−0.723	−0.421	0.203
MAD	0.173	2.140	0.057	0.051	0.046	6.842	0.030
CV	0.157	0.714	0.144	0.344	0.233	0.389	0.258
CS	45.678	0.057	−300.197	4842.189	−1532.780	0.000	−2563.660
Percentiles							
10th	1.145	1.475	0.371	0.152	0.166	11.871	0.103
25th	1.204	1.762	0.403	0.191	0.205	15.801	0.131
50th	1.309	2.229	0.473	0.200	0.250	20.541	0.162
75th	1.519	5.295	0.498	0.236	0.287	26.995	0.169
90th	1.609	7.451	0.534	0.306	0.316	31.812	0.183

Source: Standard & Poor's Compustat, Wharton Research Data Services (WRDS)

$\bar{x} = \sum_{i=1}^{20} x_i/20$. For the current ratio, for example, the ratios for all the 20 years are added and divided by 20. The means for the other ratios are calculated in a similar manner.

The second measure of central tendency presented is the median, the middle observation of an ordered set of observations. In this example, there are 20 observations, so the median is the 10th observation of the current ratio. As noted previously, the median is not affected by extreme observations, because it takes into consideration only the middle observations. In contrast, the mean uses all the observations. The mean and the median are usually close if the data contain no outliers, but if there *are* extreme observations, the mean and median may differ substantially.

The measures of variability presented are MAD, variance, and standard deviation (STD). The variance measures the average squared deviation from the mean, and the standard deviation is the square root of the variance. The MAD is the average absolute deviation from the mean. The difference between the variance and the MAD is that we square the deviations from the mean when calculating the variance and use the absolute value of the deviations when calculating the MAD.

The boundaries for the 10th, 25th, 50th, 75th, and 90th percentiles are presented to suggest the rank of the data. Recall that the percentiles indicate the percentages of observations below a certain score. For example, the boundary for the 25th percentile for JNJ's current ratio (CA/CL) is 1.6135. This means that 25 % of JNJ's current ratios were below 1.6135.

The skewness coefficient enables us to measure the shape of the distribution. A positive coefficient means that the distribution is skewed to the right, and a negative value indicates left skewness.

These statistics can be used to compare the performance of one firm to that of other firms. A high mean and a high median for the current ratio, net profit margin, and return on total assets indicate that the firm is performing well. A high current ratio (current assets divided by current liabilities) means that the firm has enough liquidity to meet current obligations, such as accounts payable and wages payable. The net profit margin measures the percentage of each dollar that goes to profits. High profit margins are a sign of profitability. The return on total assets is another measure of profitability; specifically, it measures how effectively the total assets are being utilized.

A high inventory turnover ratio is good up to a point, but very high ratios indicate that the firm may be running out of certain items in stock and losing sales to competition. Thus, inventory levels must be reasonable for the firm to maintain profitability. Generally, a high P/E ratio is also good because it means that investors believe that the firm has good growth opportunities; however, a firm may have a very high P/E ratio (approaching infinity) because of low earnings.

The total debt/total assets ratio (TD/TA in Table 4.18) is a measure of the relative amount of debt the firm has assumed. A high ratio indicates that the firm may be taking on too much debt and could be a credit risk.

The debt that a firm takes on is related to financial activity. In periods of recession, the firm is unlikely to invest in new plant and equipment, because if it did so, its existing plants and equipment would be underutilized. The firm is more likely to incur debt and invest in new projects when the economy is expanding and its plant is operating at full capacity.

Current Ratio

To determine which of the three firms was in the best liquidity position during the 20-year period, let us examine the mean and median of each firm's current ratio. JNJ had the highest current ratios, with a mean of 1.7806 and a median of 1.7745.

Figure 2.17 in Appendix 3 of Chap. 2 shows the current ratio of JNJ over time. During this period, JNJ's current ratios were close to 1.6 in the early years but later increased around 2000 to a point of 2.295 in 2001. Not following JNJ's trend, MRK had its own variance in its current ratios, remaining anyway between 0.93 and 1.4 from 1990 to 2000. Following this, it remained comfortable around 1.2 with slight variations throughout the years.

To examine the variability of the current ratio, let us look at the variance, standard deviation, coefficient of variation, and MAD. Ideally, firms like to have a high current ratio with low variability so that managers can plan from year to year. Merck had the lowest variance, 0.0451. The MAD follows the same pattern. However, it would be a mistake to conclude that JNJ's current ratio is less volatile merely on the basis of the variance and MAD because JNJ may have a lower mean and therefore a lower variance.

To eliminate this scaling problem, we use the coefficient of variation, which is calculated by dividing the standard deviation by the mean. By this standard, Merck had the lowest variability, .1732. On the basis of the central tendency and dispersion statistics, we can conclude that JNJ has experienced higher current ratios but also greater variability. While the higher ratios are beneficial, the greater dispersion is not welcomed by managers because it makes planning difficult.

The skewness coefficient can give useful information on the shape of the distribution. In this case, managers would like to have negatively or left skewed distributions because this indicates that most of the current ratios have appeared on the high end of the distribution. JNJ's skewness coefficient is 0.525, while Merck's coefficient is 0.438.

The percentiles give the rank of the data. Again, recall that the values listed for each percentile show the percentage of observations *below* that value. For example, the 10th percentile for MRK's current ratio is 1.144 indicating that 10 % of the current ratios for this 20 year period lie below 1.114. Likewise, the 10th percentile for JNJ's current ratio is 1.495. By comparing the percentiles of MRK and JNJ's current ratios, we can get a feeling of how the current ratios of these two companies have fluctuated over time. This measure shows that JNJ enjoyed superior ratios because the firm had the highest values for each of the percentiles.

Inventory Turnover Ratio

The second ratio examined is the inventory turnover ratio, which is calculated by dividing the cost of goods sold by the average value of inventory. A high inventory turnover ratio is a sign of efficiency. Both firms had similar median inventory ratios, with JNJ's being 2.496 and Merck's being 2.2285. JNJ's mean inventory ratio is 2.659 and Merck's is 3.589, showing that their ratios have a slight difference. With Merck's mean a point higher than its median, this indicates a positive skewness in the data points. JNJ's mean and median inventory ratio are very similar, showing

little skewness in the data. From these central tendency measures, we can conclude that JNJ has been more consistent than Merck with its inventory turnovers.

JNJ had the lowest variance, standard deviation, MAD, and coefficient of variation for this ratio, indicating that its inventory turnover has been less volatile than that of the other firms. A stable inventory turnover ratio makes it easier for managers to plan inventory levels. Merck performed worse compared to JNJ in this instance.

Managers also welcome a negatively skewed distribution because most of the observations are on the high end. Under this criterion, the firms did not fare well because each firm had a positively skewed distribution.

Total Debt/Total Asset Ratio

Let's look more closely at the total debt/total assets ratio (TD/TA). Both JNJ and Merck have relied steadily on debt over the 20-year time period with mean and median ratios in the 40 % range. Both firms remained somewhat constant with slight variations in their debt to asset ratios, ranging anywhere from .35 to .55, but none ever crossing the .557 mark. Both companies had similar means with JNJ and MRK's means being .448 and .455, respectively. Their medians were also, respectively, .448 and .4725. With these similarities in ratios, we can see that both Johnson & Johnson and Merck had similar debt structures in their industries.

We can see that both company's debt ratios have fluctuated very slightly, with JNJ's variance of .0031 and Merck's variance .0043. There coefficients of variation are also similar, with JNJ's being .1246 and Merck's at .1443. Since Merck's debt ratio CV is higher, it is implied that Merck's debt ratio has greater variability, even after we adjust for the mean of each company's debt ratio. Merck's superior position can be seen in the percentiles because the firm has the lowest values in all of the percentiles presented.

Net Profit Margin

The net profit margin is one of the most important ratios for managers and shareholders because it reveals what percentage of each sales dollar goes to profits. This ratio, calculated by dividing net income by sales (NI/SAL in Table 4.18), is one measure of a firm's overall profitability. As can be seen by the mean and median statistics, by this measure of profitability, Merck has been operating much more profitably than the other firms with a mean profit ratio of 0.22 and a median measure of .2. This means that about 22 % of Merck's income winds up as profits. As can be seen in Fig. 2.20 of Appendix 3 (Chap. 2), both JNJ and Merck had only positive net profit margins. The highest ratio was 0.470 in 2009 by Merck and the low was .075 by JNJ in 1992.

JNJ comes in second place with a mean net profit margin of .154 and a median of.1585. By examining the data, you can see that the values for JNJ have tended to be lower than Merck's, with the exception of a few years such as 2001 and 2002 in which JNJ's net profit margin was higher.

JNJ, however, had the lowest variability as measured by the variance, MAD, and coefficient of variation. The coefficient of variation for Merck is .3438, while JNJ had a CV of .23964.

The firms would like to have a negatively skewed distribution indicating that most of the observations are on the high side of the distribution. However, only JNJ showed a negative skewness of −.3051, while Merck showed a skewness of 2.104.

Return on Total Assets

The return on total assets is important to managers because it indicates how effectively the firm is using assets. This ratio is calculated by dividing net income by total assets (NI/TA). The higher the ratio, the greater the income generated from each dollar of total assets. As in the net profit ratio case, Merck outpaced its competitors with a mean of .2437 and a median of .2495, followed by JNJ with mean and median of .1416 and .145, respectively. JNJ had the lowest variance at .000387, with Merck's being .003213.

Price/Earnings (P/E) Ratios

Let us turn now to the price/earnings ratio. For both companies, their mean and median P/E ratios are similar. JNJ's mean P/E for this period stood at 23.2077, while Merck's was 21.5906. Their median P/E ratios, for JNJ and Merck, respectively, are 22.512 and 20.541. Because the mean is higher than the median in both cases, these results are slightly positively skewed, though not very skewed in nature due to the similarity between the companies' means and medians.

Payout Ratio

The last ratio presented is the payout ratio, which reveals what percentage of the firms' earnings are paid out in dividends. This ratio is calculated by dividing dividends per share by earnings per share. For obvious reasons, investors pay a great deal of attention to this ratio. JNJ's investors enjoyed the highest payout ratio with a mean of .3849 and a median of .366. In other words, JNJ paid out on average about 38.49 % of earnings in dividends. Merck was next with a mean of .151 and a median of .162.

From the foregoing discussion, we can conclude that, of the two pharmaceutical companies, Merck was in the best financial condition for the period 1990–2009. Although Merck didn't have the highest mean current ratio, it did have the highest mean net profit margin and return on total assets. Other things standing, like JNJ's and Merck's similar debt ratios, Merck came out financially stronger with JNJ not far behind.

Part II
Probability and Important Distributions

Chapter 5 introduces and explains such basic probability concepts as conditional, marginal, and joint probabilities. Chapter 6 discusses the analysis of discrete random variables and their distributions. Chapter 7 deals with the normal and lognormal distributions and their applications. In Chapter 8, we introduce sampling and sampling distributions. Other important continuous distributions are discussed in Chapter 9.

The examples and applications presented in Part II involve determining commercial lending rates, finding the expected value of stock price by means of the decision tree approach, determining option value via binomial and normal distributions and stock rates of return distribution, and studying auditing sampling. Other business decision applications also are explored.

Chapter 5
Probability Concepts and Their Analysis

Chapter Outline

Key Terms

Random experiment	Intersection
Basic outcomes	Addition rule
Sample points	Partition
Subset	Complement
Event	Combinatorial mathematics
Occur	Conditional probability
Simple event	Multiplication rule of probability
Basic event	Joint probability
Sample space	Marginal probability
Venn diagram	Unconditional probability
Mutually exclusive events	Simple probability
Probability	Independent
A priori probability	Bayes' theorem
Subjective probability	Prior probability
Simple event	Posterior (revised) probability
Composite event	Number of permutations
Compound event	Number of combinations
Union	Outcome trees

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_5,

5.1 Introduction

In Part I of this book, we discussed the use of descriptive statistics, which is concerned mainly with organizing and describing a set of sample measurements via graphical and numerical descriptive methods. We now begin to consider the problem of making inferences about a population from sample data. Probability and the theory that surrounds it are discussed in this chapter. These topics provide an essential foundation for the methods of making inferences about a population on the basis of a sample. A well-known example is the election poll, in which pollsters select at random a small number of voters to question in order to predict the winner of an election. Probability is also used in daily decision making. For example, investment decisions are based on the investor's assessment of the probable future returns of various investment opportunities, and such assessments are often based on some sample information.

In this chapter, we first discuss how basic concepts such as random experiment, outcomes, sample space, and event can be used to analyze probability. Then, we investigate alternative events and their probabilities. In probability theory, conditional, joint, and marginal probabilities are the most important concepts in analyzing statistical business and economic problems. Therefore, they are explored in detail in this chapter. We also discuss independent and dependent events and Bayes' theorem. Finally, four business applications of probability are demonstrated.

5.2 Random Experiment, Outcomes, Sample Space, Event, and Probability

A *random experiment* is a process that has at least two possible outcomes and is characterized by uncertainty as to which will occur.

Each of the following examples involves a random experiment:

1. A die is rolled.
2. A voter is asked which of four candidates he or she prefers.
3. A person is asked whether President Bush should order US troops to liberate Kuwait.
4. The daily change in the price of silver per ounce is observed.

When a die is rolled, the set of basic outcomes comprises 1 through 6; these basic outcomes represent the various possibilities that can occur. In other words, the possible outcomes of a random experiment are called the *basic outcomes*. The set of all basic outcomes is called the *sample space*. Thus, basic outcomes are equivalent to sample points in a sample space.

Suppose you are interested in getting an even number in rolling a die; in this case, the event is rolling a 2, 4, or 6, which is a *subset* of {1, 2, 3. 4, 5, 6}. In other words, an *event* is a set of basic outcomes from the sample space, and it is said to

occur if the random experiment gives rise to one of its constituent basic outcomes. Each basic outcome within each event (e.g., {2} {4} {6}) can also be called a *simple event*. Hence, an event is a collection of one or more simple events. Finally, a *basic event* is a subset of the sample space. The concepts of random experiment, outcomes, sample space, and event, then, are fundamental to an understanding of probability.

5.2.1 Properties of Random Experiments

The starting point of probability is the random experiment. Random experiments have three properties:

1. They can be repeated physically or conceptually.
2. The set consisting of all of possible outcomes—that is, the sample space—can be specified in advance.
3. Various repetitions do not always yield the same outcome.

Simple examples of conducting a random experiment include rolling dice, tossing a coin, and drawing a card from a deck of 52 playing cards.

Because of uncertainty in the business environment, business decision making is a tricky and an important skill. If the executive knew the exact outcomes of the courses of action available, he or she would have no difficulty making optimal decisions. However, the executive generally does not know the exact outcome of a decision. Thus, business executives spend much time evaluating the probabilities of various alternative outcomes. For example, an executive may need to determine the probability of extensive employee turnover if the firm moves to another area. Or a business decision maker may want to evaluate the impact of changes in economic indicators such as interest rate, inflation, and gross national product (GNP) on a company's future earnings.

5.2.2 Sample Space of an Experiment and the Venn Diagram

For convenience, we can represent each outcome of a random experiment by a set of symbols. The symbol S is used to represent the *sample space* of the experiment. As we have noted, the sample space is the set of all basic outcomes (simple events) of the random experiment. In the foregoing die-rolling example, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. When a person takes a driver's license test, the sample space contains only two elements: $S = \{P, F\}$, where P indicates a pass and F a failure. In a stock price forecast, the sample space could contain three elements: $S = \{U, D, N\}$, where U, D, and N represent movement up, movement down, and no change in the price of a stock. In sum, the different basic outcomes of an experiment are often referred to as *sample points* (simple events), and the set of all possible outcomes is called the *sample space*. Thus, the sample points (simple events) form the sample space.

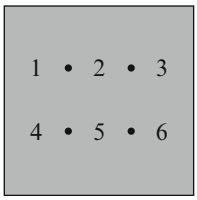


Fig. 5.1 Venn diagram showing six different sample points

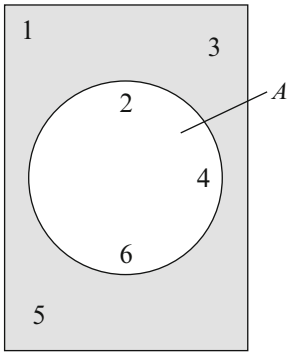


Fig. 5.2 Venn diagram for event *A*

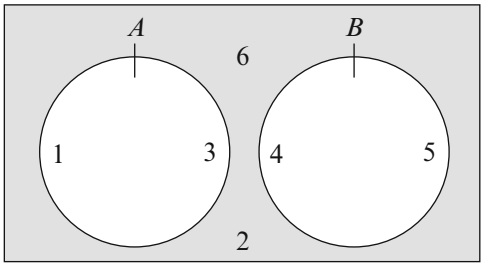


Fig. 5.3 Venn diagram for mutually exclusive events

A *Venn diagram* can be used to describe graphically various basic outcomes (simple events) in a sample space. The rectangle represents the sample space, and the points are basic outcomes. Events are usually represented by circles or rectangles. Figure 5.1 shows a Venn diagram. The elements labeled represent the six basic outcomes of rolling a die. In Fig. 5.2, the circle indicates the event of all even numbers that can result from rolling a single die. Let event $A = \{2, 4, 6\}$. Again, the sample space is the possible outcomes of rolling a die. Figure 5.3 shows events $A = \{1, 3\}$ and $B = \{4, 5\}$. When two events have no basic outcome in common, they are said to be *mutually exclusive events*. When events have some elements in common, the *intersection* of the events is the event that consists of the common elements. Say we have one event $A = \{2, 3, 4, 6\}$ and another event $B = \{2, 3, 5\}$. The intersection of these events is shown in Fig. 5.4. The common elements are 2 and 3.

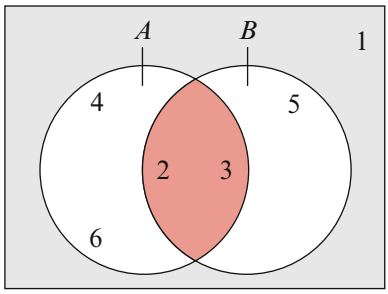


Fig. 5.4 Venn diagram for the intersection of events *A* and *B*

5.2.3 *Probabilities of Outcomes*

The *probability* of an event is a real number on a scale from 0 to 1 that measures the likelihood of the event's occurring. If an outcome (or event) has a probability of 0, then its occurrence is impossible; if an outcome (or event) has a probability of 1.0, then its occurrence is certain. Getting *either* a head *or* a tail in a coin toss is an example of an event that has a probability of 1.0. Because there are only two possibilities, either one event or the other is certain to occur. An event with a zero probability is an impossible event, such as getting both a head and a tail when tossing a coin once.

When we roll a fair die, we are just as likely to obtain any face of the die as any other. Because there are six faces to a die, we generally say the "outcome" of the toss can be one of six numbers: 1, 2, 3, 4, 5, 6.

The probability of an outcome can be calculated by the classical approach, the relative frequency approach, or the subjective approach. The first two approaches are discussed in this section, the third approach in the next.

Classical probability is often called *a priori probability*, because if we keep using orderly examples, such as fair coins and unbiased dice, we can state the answer in advance (a priori) without tossing a coin or rolling a die. In other words, we can make statements based on logical reasoning before any experiments take place. Classical probability defines the probability that an event will occur as

$$\text{Probability of an event} = \frac{\text{number of outcomes contained in the event}}{\text{total number of possible outcomes}} \tag{5.1}$$

Note that this approach is applicable only when all basic outcomes in the sample space are equally probable.

For example, the probability of getting a tail upon tossing a fair coin is

$$P(\text{tail}) = \frac{1}{1+1} = \frac{1}{2}$$

And for the die-rolling example, the probability of obtaining the face 4 is

$$P(4) = \frac{1}{6}$$

The relative frequency approach to calculating probability requires the random experiment to take place as defined in Eq. 5.2:

$$P(o = e_i) = \frac{n_i}{N} \qquad or \qquad P(e_i) = \frac{n_i}{N} \tag{5.2}$$

where

o = outcome
e_i = outcome associated with ith event
n_i = number of times the ith outcome occurs
N = total number of times the trial is repeated

From Eq. 5.2, we know that we can obtain the relative frequency by dividing the total number of trials being repeated into the number of ith outcomes occurring. Another explanation for Eq. 5.2 would be $P(\mathrm{e}_i) = f_i/N$, where f_i equals the number of favorable outcomes for event e_i and N equals the total outcome in sample space, S.

The credit cards issued by Citicorp in 1984 are listed here (the data is from *Fortune* magazine, February 4, 1985, page 21):

Credit card	Number of cards issued (in millions)
Visa and MasterCard	6.0
Diners club	2.2
Carte blanche	.3
Choice	1.0

Visa and MasterCard credit cards are issued by thousands of banks, including Citicorp. The other three credit cards listed above are issued by Citicorp only. If one Citicorp credit card customer is selected randomly, the probability that the customer selected uses one of Citicorp's own credit cards is

$$\frac{2.2 + .3 + 1.0}{2.2 + .3 + 1.0 + 6.0} = .368.$$

Example 5.1 Toss a Fair Coin. Suppose a random experiment is to be carried out and we are interested in the chance of occurrence of a particular event. The concept of probability can help us, because it provides a numerical measure for the likelihood of an event or a set of events occurring.

We conduct 50 experiments of tossing a fair coin in different sample sizes ($N = 10, 20, \ldots, 500$). The results of these experiments are presented in Table 5.1. In Table 5.1, the first column represents that the ith ($i = 1, 2, \ldots, 50$) experiment has been done; the second column, N, shows the number of times a coin has been

Table 5.1 Frequency and proportion in tossing a fair coin

MTB > PRINT C51–C53			
Data display			
Row	N	f	p
1	10	7	0.700000
2	20	12	0.600000
3	30	17	0.566667
4	40	19	0.475000
5	50	22	0.440000
6	60	34	0.566667
7	70	30	0.428571
8	80	44	0.550000
9	90	52	0.577778
10	100	37	0.370000
11	110	57	0.518182
12	120	64	0.533333
13	130	59	0.453846
14	140	63	0.450000
15	150	78	0.520000
16	160	78	0.487500
17	170	95	0.558824
18	180	88	0.488889
19	190	90	0.473684
20	200	100	0.500000
21	210	117	0.557143
22	220	109	0.495455
23	230	115	0.500000
24	240	119	0.495833
25	250	130	0.520000
26	260	128	0.492308
27	270	137	0.507407
28	280	126	0.450000
29	290	140	0.482759
30	300	158	0.526667
31	310	161	0.519355
32	320	157	0.490625
33	330	160	0.484848
34	340	165	0.485294
35	350	168	0.480000
36	360	188	0.522222
37	370	201	0.543243
38	380	199	0.523684
39	390	190	0.487179
40	400	198	0.495000
41	410	220	0.536585
42	420	192	0.457143
43	430	207	0.481395

(continued)

Table 5.1 (continued)

MTB > PRINT C51–C53			
Data display			
Row	N	f	p
44	440	210	0.477273
45	450	226	0.502222
46	460	217	0.471739
47	470	239	0.508511
48	480	243	0.506250
49	490	233	0.475510
50	500	243	0.486000

tossed for experiment; the third column lists the number of times heads appeared; and the fourth column gives the proportion of heads that appeared.[1] Figure 5.5 is the figure generated by MINITAB in terms of the data given in the fourth column.

In this example, we know that if the coin is fair, anytime we toss the coin, the probability of our getting heads is $\frac{1}{2}$. One important property of probability is that *the sum of the probabilities of all outcomes must be equal to 1*. The sum of the probabilities must equal 1 because the possible outcomes are collectively exhaustive and mutually exclusive. From our previous example of the toss of a coin, the outcomes are mutually exclusive and collectively exhaustive because we have included all the possible basic outcomes (simple events) that can occur. Mathematically, we can express this property for our roll of the die as

$$P(H) + P(T) = 1 \tag{5.3}$$

where $P(H)$ and $P(T)$ represent the probability of getting heads and that of getting tails, respectively.

Note that the results of Table 5.1 are obtained by tossing a coin again and again. Tossing it only 4 (or even 20) times would not be enough to average out the chance fluctuations shown in Table 5.1. When N is large enough, however, the relative frequency of tossing a coin for heads moves toward $\frac{1}{2}$, as indicated in Fig. 5.5.

So far, we have used both the classical and the relative frequency approaches to define probability. We now summarize the definition of probability in terms of relative frequency. Let n_i be the number of occurrences of event i in N repeated trials.

Then under the relative frequency concept of probability, the probability that event i will occur is the relative frequency (the ratio n_i/N) as the number of trials N becomes infinitely large. Alternatively, the probability can be interpreted as the proportion of times the ith event (n_i/N) occurs in the long run (N becomes infinitely large) when conditions are stable.

[1] This set of data was generated by the Bernoulli process, which will be discussed in the next chapter.

```
MTB >  Plot 'P'* 'N';
SUBC>    Connect;
SUBC>    Axis 1;
SUBC>      Label "SAMPLE SIZE";
SUBC>    Axis 2;
SUBC>      Label "PROPORTION".
```

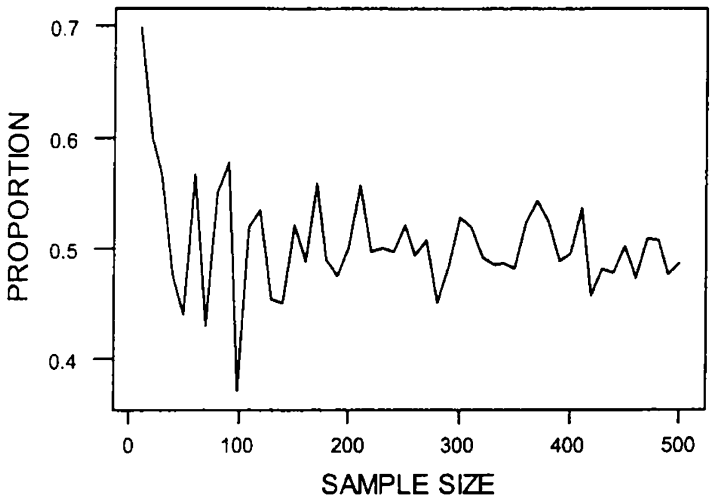


Fig. 5.5 M1NITAB output for the proportion of heads in *N* tosses of a fair coin

5.2.4 Subjective Probability

An alternative view about probability, which does not depend on the concept of repeatable random experiments, defines probability in terms of a subjective, or personalistic, concept. According to this concept of *subjective probability*, the probability of an event is the degree of belief, or degree of confidence, an individual places in the occurrence of an event on the basis of whatever evidence is available. This evidence may be data on the relative frequency of past occurrences, or it may be just an educated guess. The individual may assign an event the probability of 1, 0, or any other number between those two. Here are a few examples of situations that require a subjective probability:

1. An individual consumer assigns a probability to the event of purchasing a TV during the next quarter.
2. A quality control manager asserts the probability that a future incoming shipment will have 1.5 % or fewer defective items.
3. An auditing firm wishes to determine the probability that an audited voucher will contain an error.

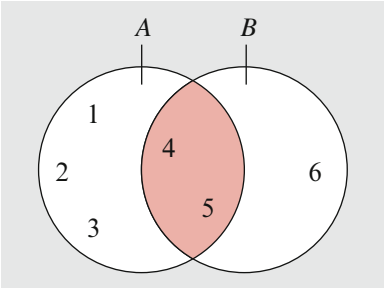


Fig. 5.6 Venn diagram for the intersection of events *A* and *B*

4. An investor ponders the probability that the Dow Jones closing index will be below 3,000 at some time during a 3-month period beginning on November 10, 1992.

5.3 Alternative Events and Their Probabilities

As we have stated, an event is the result of a random experiment consisting of one or more basic outcomes. If an event consists of only one basic outcome, it is a *simple event*; if it consists of more than one basic outcome, it is a *composite event*. In the die-rolling experiment discussed in Fig. 5.1, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$.

Suppose we are interested in the event E, where the outcome is 1 or 6. We can clearly describe the event E as $E = \{1.\ 6\}$. An event E is a subset of the sample space S. This is a composite event because it includes the simple events $\{1\}$ and $\{6\}$. The subset definition enables us to define an event in general.

In the tossing of a fair die, suppose that event A represents the faces 1, 2, 3, 4, and 5 and event B the faces of 4, 5, and 6. Graphically, the relationship between basic outcomes and events can be represented as shown in Fig. 5.6. The intersection of these two events is the faces 4 and 5, because these faces are common to both events.

5.3.1 Probabilities of Union and Intersection of Events

An event can often be viewed as a composite of two or more other events. Such an event, called a *compound event*, can be classified as union or as intersection. The *union* of two events A and B is the event that occurs when either A or B or both occur on a single performance of the experiment. For example, if event B is getting an even number (2, 4, or 6) on a die toss and event A is getting a number 1 or 2, then the union of events A and B, which we represent as $A \cup B$, is 1, 2, 4, and 6. The union $A \cup B$ is indicated in Fig. 5.7.

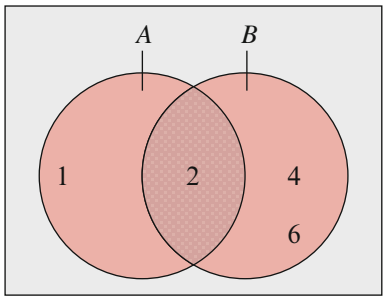


Fig. 5.7 Venn diagram for $A \cup B$

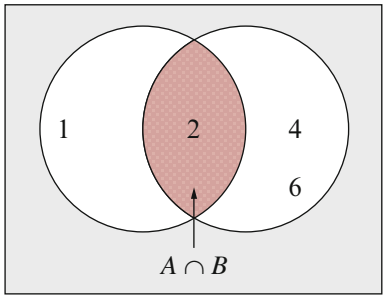


Fig. 5.8 Venn diagram for $A \cap B$

The *intersection* of two events A and B is the event that occurs when both A and B occur on a single performance of the experiment. That is, the common members make up the intersection of two events. Because 2 is the only common sample point in our two events, the intersection of events A and B, which we represent as $A \cap B$, is 2. This intersection is indicated by the shaded area in Fig. 5.8.

Example 5.2 Pick a Card, Any Card. To illustrate the union and intersection of events, we shall use a standard deck of cards. We know that there are 52 cards in a deck (13 spades, 13 hearts, 13 diamonds, and 13 clubs) and 4 cards for each number (see Table 5.2). Using this information, let's calculate the probability of the union and intersection of events in the sample space of a deck of playing cards.

5.3.1.1 Probability of Union

To assess the probability of union, first, imagine we randomly select one card from the deck. Let event $A = \{\text{club}\}$ and event $B = \{\text{heart or diamond}\}$. Let $A \cup B$ denote the union, so $A \cup B = \{\text{club, heart, diamond}\}$.

The union of A and B means the event "A or B" occurs. We can now compute the mathematical probability of A or B:

$$P(A) = \frac{13}{52} = \frac{1}{4} \quad \text{and} \quad P(B) = \frac{13 + 13}{52} = \frac{1}{2}$$

Table 5.2 Playing card sample space

Spades	Hearts	Diamonds	Clubs
A	A	A	A
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9
10	10	10	10
J	J	J	J
Q	Q	Q	Q
K	K	K	K

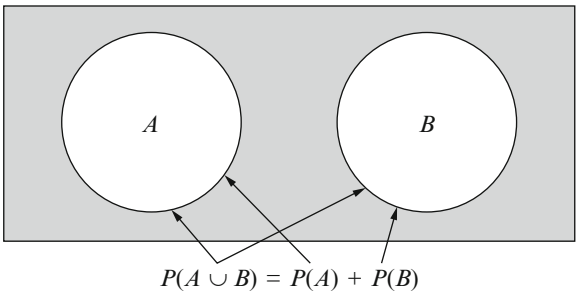


Fig. 5.9 Venn diagram for the probability of two mutually exclusive events

The probability of getting a club, a heart, or a diamond is obtained by adding the number of club, heart, and diamond cards and dividing by the total number of cards, 52. As a result, the probability of drawing a card that is a member of the union of these two events is

$$P(A \cup B) = P(A) + P(B) = \tfrac{1}{4} + \tfrac{1}{2} = \tfrac{3}{4}$$

Thus, we have a $\frac{3}{4} = 75\ \%$ chance of randomly drawing a single card that is a club *or* a heart *or* a diamond.

If A and B are mutually exclusive, the probability formula for a union of A and B is

$$P(A \cup B) = P(A) + P(B) \tag{5.4}$$

The rule for obtaining the probability of the union of A and B as indicated in Eq. 5.4 is the *addition rule* for two events that are mutually exclusive. This addition rule is illustrated by the Venn diagram in Fig. 5.9, where we note that the area of two circles taken together (denoting $A \cup B$) is the sum of the areas of the two circles.

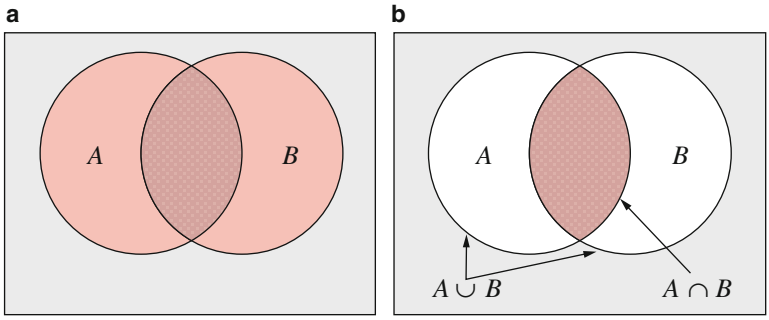


Fig. 5.10 Venn diagram for the probability of two events that are not mutually exclusive: (**a**) $A \cup B$ and (**b**) $A \cap B$

As another example, if $A =$ all clubs and $B =$ all diamonds, then

$$P(A \cup B) = \tfrac{13}{52} + \tfrac{13}{52} = \tfrac{26}{52} = \tfrac{1}{2}$$

A new pharmaceutical product is about to be introduced commercially, and both Upjohn and Merck want to be the first to put the product on the market. An industrial analyst believes that the probability is .40 that Upjohn will be first and .25 that Merck will be first. If the analyst's beliefs are correct, what is the probability that either Upjohn or Merck will be first, assuming that a tie does not occur?

Let A be the event that Upjohn is first. Let B be the event that Merck is first. Then, from Eq. 5.4, we have $P(A \cup B) = .40 + .25 = .65$. Consequently, the probability that either Upjohn or Merck will be first is .65.

If A and B are not mutually exclusive, then the simple probability of union defined in Eq. 5.4 must be modified to take the intersection into account and thereby avoid double counting:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{5.5}$$

The rule for obtaining the probability of the union of A and B as indicated in Eq. 5.5 is the addition rule for two events that are *not* mutually exclusive. This addition rule is illustrated by Fig. 5.10. In Fig. 5.10a, the event $A \cup B$ is the sum of the areas of circles A and B. The event $A \cap B$ is the shaded area in the middle, as indicated in Fig. 5.10b. When we add the areas of circles A and B, we count the shaded area twice, so we must subtract it to make sure it is counted only once.

If, instead, $A =$ all diamonds and $B =$ all diamonds or all hearts, then

$$P(A \cup B) = \tfrac{1}{4} + \tfrac{1}{2} - \tfrac{1}{4} = \tfrac{1}{2}$$

Midlantic Bank in New Jersey gives summer jobs to two Rutgers University business school students, Mary Smith and Alice Wang. The bank personnel

Table 5.3 Family size data

Number of children	0	1	2	3	4	5	6	7 or more
Proportion of families having this many children	.04	.11	.29	.26	.14	.10	.05	.01

manager hopes that at least one of these students will decide to work for the bank upon graduation. Assume that the probability that Mary will decide to work for the bank is .4, the probability for Alice is .3, and the probability that both will decide to work for the bank is .2. Then, the probability that the personnel manager's hopes will be fulfilled is

$$P(A \cup B) = .4 + .3 - .2 = .5.$$

Example 5.3 Probability Analysis of Family Size. Table 5.3 contains data on the size of families in a certain town in the United States in 1992. If we randomly choose a family from this town, what is the probability that this family includes three or more children?

Using Eq. 5.4, we can calculate the answer as

$$\begin{aligned} P(3,\ 4,\ 5,\ 6 \text{ or more}) &= P(3) + P(4) + P(5) + P(6) + P(7 \text{ or more}) \\ &= .26 + .14 + .10 + .05 + .01 \\ &= .56 \end{aligned}$$

5.3.1.2 Probability of Intersection

If $A = \{\text{diamond}\}$ and $B = \{\text{diamond or heart}\}$, then $A \cap B = \{\text{diamond}\}$ = set of points that are in both A and B. Using Table 5.2, we obtain

$$\begin{aligned} P(A) &= \tfrac{13}{52} = \tfrac{1}{4} \\ P(B) &= (13 + 13)/52 = \tfrac{1}{2} \\ P(A \cap B) &= \tfrac{13}{52} = \tfrac{1}{4} \end{aligned}$$

Thus, the probability of drawing a diamond *and* drawing a diamond or a heart is the probability of drawing a diamond, which is $\frac{1}{4}$, or 25 %.

From Eq. 5.5, we can define the probability of an intersection as

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) \tag{5.6}$$

If, instead, A = all diamonds and B = all diamonds or all hearts, then

$$P(A \cap B) = \tfrac{1}{4} + \tfrac{1}{2} - \tfrac{1}{2} = \tfrac{1}{4}$$

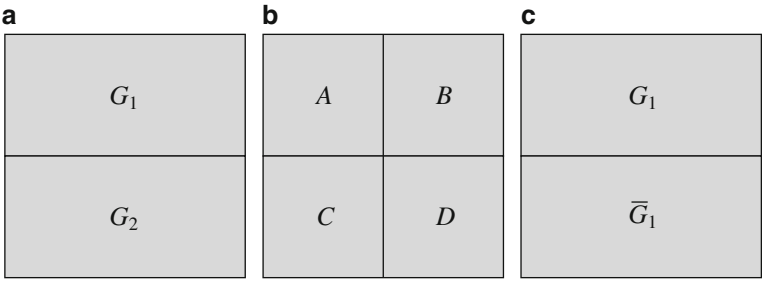


Fig. 5.11 Venn diagrams of mutually exclusive events, showing partitions and complements

5.3.2 *Partitions, Complements, and Probability of Complements*

Now, suppose we randomly choose a card from the deck. Let $A =$ (red suit) and $B =$ (black suit). A card cannot be a member of both a red suit *and a* black suit. Therefore, we say A and B are *mutually exclusive events*; they have no basic outcomes in common. In addition, if mutually exclusive events A and B cover the whole sample space S, we call the collection of events A and B a *partition* of S. Another alternative is to partition the card deck sample space as follows:

$$A = \{\text{club}\} \quad B = \{\text{diamond}\} \quad C = \{\text{heart}\}, \quad D = \{\text{spade}\}$$

Those four events—A, B, C, and D—are mutually exclusive and collectively exhaustive; the collection of these events is called a partition of sample space S, which can be explicitly defined as

$$S = \{A, B, C, D\} \tag{5.7}$$

Equation 5.7 can itself be partitioned again into G_1 and G_2 as

$$S = \{G_1, G_2\}, \tag{5.8}$$

where $G_1 = \{A, B\}$ and $G_1 = \{C, D\}$. G_1 consists of exactly those cards that are not in G_2. We therefore call G_2 the *complement* of G_1, denoted by $\bar{G}_1$ (which is read "not G_1"). In other words, $\bar{G}_1$ represents a set of cards that are not in G_1: $\bar{G}_1 = \{C, D\}$.

Figure 5.11 depicts three different types of partitions. Figure 5.11a depicts two mutually exclusive sets, G_1 and G_2. Figure 5.11b shows mutually exclusive events, A, B, C, and D. $\bar{G}_1$ is the complement of G_1 in Fig. 5.11c. $\bar{G}_1$ and G_1 are mutually exclusive for a simple partition.

Table 5.4 Employed workers in 1987

Occupation	Relative frequency
Male worker	.552
Managerial/professional	.137
Technical/sales/administrative	.110
Service	.053
Precision production, craft, and repair	.110
Operators/fabricators	.115
Farming, forestry, and fishing	.027
Female worker	.448
Managerial/professional	.109
Technical/sales/administrative	.202
Service	.081
Precision production, craft, and repair	.010
Operators/fabricators	.040
Farming, forestry, and fishing	.006

Source: Statistical Abstract of the United States: 1989, p. 388

5.3.2.1 Probability of Complement

Because an event and its complement, $\{E, \overline{E}\}$, constitute a simple partition, these events are mutually exclusive. By Eq. 5.4,

$$P(E \cup \overline{E}) = P(E) + P(\overline{E}) \tag{5.9}$$

E and $\overline{E}$ constitute all of the sample space, so

$$P(E \cup \overline{E}) = 1 \tag{5.10}$$

Substituting Eq. 5.10 into Eq. 5.9, we obtain

$$\begin{aligned} 1 &= P(E) + P(\overline{E}) \\ P(\overline{E}) &= 1 - P(E) \end{aligned} \tag{5.11}$$

Recalling Eq. 5.8 about playing cards, where G_1 represents a club or a diamond, we find that the probability of $\bar{G}_1$ (neither a club nor a diamond) is

$$\begin{aligned} P(\bar{G}_1) &= 1 - P(G_1) \\ &= 1 - 26/52 \\ &= 1/2 \end{aligned}$$

In 1987, 112,440,000 workers were employed in the United States. Table 5.4 shows the relative frequencies of these employed workers, classified by different types of occupations.

If we need to select a worker randomly from the population and determine his or her occupation, the probability that the worker will not be in a technical/sales/administrative occupation can be calculated as follows:

P(nontechnical/sales/administrative occupation)
$= 1 - P$(technical, sales, or administrative occupation)
$= 1 - P$(male, technical, sales, or administrative occupation) $- P$(female, technical, sales, or administrative occupation)
$= 1 - .110 - .202 = .688$.

5.3.3 *Using Combinatorial Mathematics to Determine the Number of Simple Events*

The purpose of introducing combinatorial mathematics here is to show how the number of simple events can be determined and the probability computed. *Combinatorial mathematics* is the mathematics that develops counting principles and techniques in terms of permutations and combinations, which are discussed in Appendix 1. For example, a simple rule for finding the number of different samples of r auto part items selected from n auto part items in doing quality control sampling can be derived from the combination formula discussed in this section. According to the combination formula developed in Appendix 1, the total number of possible combinations of samples is

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{5.12}$$

where n is the number of possible objects (items), r is the number of objects to be selected, and the factorial symbol (!) means that, say,

$$n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$$
$$(n-r)! = (n-r)(n-r-1)\cdots 3 \cdot 2 \cdot 1$$

For example, $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$. (The quantity of 0! is defined as equal to 1.)

Example 5.4 Possible Combinations in Selecting Gifts. The United Jersey Bank in New Jersey is giving out gifts to depositors. If eligible, depositors may choose any two out of six gifts. How many possible combinations of gifts can different depositors select? This question can be answered either manually or by combinatorial mathematics.

Manual Method

Let g_1, g_2, g_3, g_4, g_5, and g_6 represent first gift, second gift, third gift, fourth gift, fifth gift, and sixth gift. The number of possible combinations of two gifts chosen from among these six gifts is 15:

g_1, g_2	g_2, g_3	g_3, g_5
g_1, g_3	g_2, g_4	g_3, g_6
g_1, g_4	g_2, g_5	g_4, g_5
g_1, g_5	g_2, g_6	g_4, g_6
g_1, g_6	g_3, g_4	g_5, g_6

Combinatorial Mathematics Method

Or, if we have less paper, we can use Eq. 5.5 and find the number of possible combinations as follows:

$$\binom{6}{2} = \frac{6!}{(2!)(6-2)!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1)(4 \cdot 3 \cdot 2 \cdot 1)} = 15$$

This result agrees with the result we obtained manually.

If both n and r are large, combinatorial mathematics is the far better method for counting the number of outcomes in the sample space. Trust me.

Example 5.5 Just Take the Toaster. If the two gifts are randomly selected, what is the probability of gift 1 being selected? Well, there are five combinations that include gift 1, and there are 15 possible combinations. The probability of gift 1 being selected, then, is $P = 5/15 = 1/3$.

5.4 Conditional Probability and Its Implications

5.4.1 Basic Concept of Conditional Probability

Conditional probability is the probability that an event will occur, given that (on the condition that) some other event *has* occurred. The concept of conditional probability is relatively simple. In the example involving playing cards that was discussed in Sect. 5.3, we have 13 spades, 13 hearts, 13 diamonds, and 13 clubs. Suppose we put 13 spades on the table and then select a card randomly from that group.[2] What is the probability that the card's face value will be 2, P(S2), given that it is a spade? Here, we have changed the condition under which the experiment is performed, because we are now considering only a subset of the population: just the spades. To obtain a new probability for each element of this subpopulation, we simply find the total probability of the subpopulation (spades) and then divide the probability of each event in the subpopulation by the total probability. We know that the total probability of the subpopulation is 13/52 [(1/52)(13)]. The new probabilities we assign are

[2] This is equivalent to randomly drawing a card from the deck and finding that it is a spade.

$$P(\text{S2} \mid \text{spades}) = \frac{P(\text{S2})}{P(\text{spades})} = \frac{1/52}{13/52} = \frac{1}{13}$$
$$P(\text{S3} \mid \text{spades}) = \frac{1/52}{13/52} = \frac{1}{13}$$
$$\vdots$$
$$P(\text{SA} \mid \text{spades}) = \frac{1/52}{13/52} = \frac{1}{13}$$

where S2, S3, and SA represent the 2, 3, and ace of spades. The notation|means "given." For example, P(S2|spades) means the probability of drawing a 2 of spades, given that the card is a spade.

If we let A = the event of picking a spade from the deck and B = the card being a 2, the conditional probability of this drawing is written as

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{1/52}{13/52} = \frac{1}{13}$$

where $P(B \cup A)$ is the probability that the card is a 2 of spades and $P(A)$ is the probability that the card is a spade. Now, we can give the formula for conditional probability as

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} \tag{5.13}$$

Assume that J, Q, and K are greater than 10, as defined in Table 5.2. If we let A represent the event that the card we draw is a spade and let B represent the event that the card is a jack, queen, or king, then

$$P(B \cap A) = 1/52 + 1/52 + 1/52 = 3/52$$

$P(A)$ is the probability that the card we pick up is a spade, or $13/52 = 1/4$. Then,

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{3/52}{1/4} = \frac{3}{13} = 23.08\%$$

The conditional probability $P(B \mid A) = 3/13$ can be shown on a Venn diagram as indicated in Fig. 5.12, where $A \cap B$ takes 23.08 % of the total area of A, which means that $P(B \mid A) = 23.08$ %.

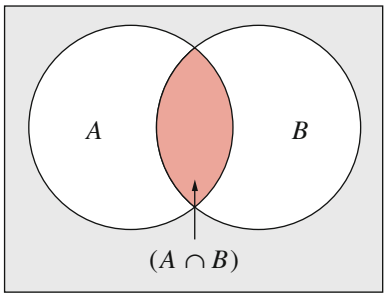


Fig. 5.12 Venn diagrams of conditional probability

5.4.2 *Multiplication Rule of Probability*

An immediate consequence of the definition of conditional probability is the *multiplication rule of probability*, which expresses the probability of an intersection in terms of the probability of an individual event and the conditional probability. It can be derived as follows. From Eq. 5.13, we know that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{5.13a}$$

From Eqs. 5.13 and 5.13a, we obtain

$$P(B \cap A) = P(B \mid A)P(A) \tag{5.14}$$

$$P(A \cap B) = P(A \mid B)P(B) \tag{5.15}$$

Clearly, $(B \cap A) = (A \cap B)$. Thus, from Eqs. 5.14 and 5.15, we obtain

$$P(A \cap B) = P(B \mid A)P(A) = P(A \mid B)P(B) \tag{5.16}$$

For example, suppose 30 % of all students receive a grade of C (event A). Of all students who receive C, 60 % are male (event B). This is a conditional probability because we are limiting ourselves to male students. In symbols, $P(B|A) = .60$. What is the probability of a randomly selected student who is a male having a grade of C?

The event of a male student with a grade of C is the intersection *of* A and B. We also know that

$$P(A) = .3$$

$$P(B|A) = .6$$

Then, from our rule of Eq. 5.16, we find that

$$P(A \cap B) = P(A)P(B|A) = (.3)(.6) = .18$$

Example 5.6 Joint Probability on Wall Street. Let A be the event that the stock market will be bullish next year, and let B be the event that the stock price of Meridian Company will increase by 10 % next year. An investment analyst would like to estimate the probability that the stock price of Meridian Company will increase *and* that the stock market will be bullish next year. Let

$$\begin{aligned} P(A) &= \text{probability that the stock market will be bullish next year} \\ &= 60\,\text{percent} \\ P(B|A) &= \text{probability that the stock price of Meridian Company will increase by 10} \\ &\quad \text{percent, given that the stock market will be bullish} \\ &= 30\,\text{percent} \end{aligned}$$

Using Eq. 5.15, we obtain

$$P(A \cap B) = (.6)(.3) = .18$$

This implies that there is about an 18 % chance that the stock price of Meridian Company will increase and that the stock market will be bullish.

The probability $P(A \cap B)$ of Eq. 5.15 is called the joint probability, which is discussed in Sect. 5.5 in further detail. Equation 5.15 can also be used to derive Bayes' theorem, which is discussed in Sect. 5.7.

5.5 Joint Probability and Marginal Probability

In this section, we will examine joint and marginal probabilities and their relationships to the conditional probability we discussed in Sect. 5.4.

5.5.1 Joint Probability

In many applications, we are interested in *joint probability*, the probability of two or more events occurring simultaneously. To illustrate joint probabilities, consider the data in Table 5.5. These figures represent the results of a market survey in which 500 persons were asked which of two competitive soft drinks they preferred, soft drink 1 from company I or soft drink 2 from company II. To simplify the discussion, as shown in Table 5.5, we use M, F, S_1, and S_2 to represent male, female, prefers

Table 5.5 500 persons classified by sex and product preference

Product preference			
Sex	S_1	S_2	Total
Male	100	160	260
Female	200	40	240
Total	300	200	500

Table 5.6 Joint probability table for 500 persons classified by sex and product preference

Product preference			
Sex	S_1	S_2	Marginal probability
Male	.20	.32	.52
Female	*.40*	*.08*	*.48*
Marginal probability	.60	.40	1.00

soft drink 1, and prefers soft drink 2. Hence, the joint outcome that an individual is both male and prefers soft drink l is denoted as "M and S_1," and the joint probability that a randomly selected individual is male and prefers soft drink 1 is $P(M \text{ and } S_1) = P(M \cap S_1)$. This probability is

$$P(M \cap S_1) = \frac{100}{500} = 0.20$$

Other joint probabilities can be calculated similarly. Table 5.6 is a joint probability table obtained by dividing all entries in Table 5.5 by the total number of individuals.

For two events, a joint probability is the probability of the intersection of two events—in other words, the probability that both events will occur at the same time. As we saw in Eq. 5.16, the joint probability can be defined as

$$P(A \cap B) = P(B \mid A)P(A) = P(A \mid B)P(B) \tag{5.16}$$

$P(A \cap B)$ can be also represented as $P(A \text{ and } B)$. It is used to denote the probability that both events A and B will occur. This equation implies that a joint probability is the product of a marginal probability [either $P(A)$ or $P(B)$] and a conditional probability [either $P(B|A)$ or $P(A|B)$]. In Table 5.6, marginal probabilities are presented in the last row and the last column. The concept of marginal probability is discussed later in this section.

In the case where we drew a spade from among 13 spades, $P(A) = 13/52$ and $P(B \mid A) = 1/13$. So the joint probability is

$$P(B \cap A) = (1/13)(13/52) = 1/52$$

Table 5.7 Eighty persons classified by sex and race

Sex	Black	White	Total
Male	5	35	40
Female	25	15	40
Total	30	50	80

5.5.2 Marginal Probabilities

In addition to joint probability, we can also obtain from Table 5.6 probabilities for the two classifications "sex" and "product preference." These probabilities, which are shown in the margins of the joint probability table, are referred to as *marginal probabilities or unconditional probabilities*. For example, the marginal probability that a randomly chosen individual is female is $P(F) = .48$, and the marginal probability that a person prefers soft drink 1 is $P(S_1) = .60$. In these cases, the marginal probabilities for each classification are obtained by summing the appropriate joint probabilities. Because marginal probability is a probability of a simple event, it is often called the *simple probability*.

Armed with this information on joint probability and marginal probability, we can calculate conditional probability as indicated in Eq. 5.13. The probability that the individual is female *and* prefers soft drink 1, for example, can be calculated, in terms of data listed in Table 5.6, as $P(S_1 \cap F) = .40$. We also know that $P(F) = .48$. Substituting this information into Eq. 5.13, we obtain

$$P(S_1 \mid F) = \frac{P(S_1 \cap F)}{P(F)} = \frac{.40}{.48} = \frac{40}{48} = \frac{5}{6}$$

For comparison, we now calculate

$$P(F \mid S_1) = \frac{P(F \cap S_1)}{P(S_1)} = \frac{.40}{.60} = \frac{2}{3}$$

Note that $P(F \mid S_1) \neq P(S_1 \mid F)$. Note that $P(S_1 \cap F)$ can be calculated by dividing 500 into 200 (see the data presented in Table 5.5).

To further illustrate this point, let's consider the following example.

Example 5.7 Classifying Students by Two Criteria. Suppose we consider the problem of randomly selecting 1 student as a representative from a class of 80 students. In this class, there are 5 black male students, 25 black female students, 35 white male students, and 15 white female students, as indicated in Table 5.7.
Let

B = event that the student is black
W = event that the student is white
M = event that the student is male
F = event that the student is female

Events B and W classify the students by race. Events M and F classify them by sex.

Because each event represents only one of the different classifications (sex or race), the probabilities of these events are called marginal probabilities. Marginal probabilities for race and sex can be calculated as

$$P(B) = \frac{30}{80}, \quad P(W) = \frac{50}{80}, \quad \text{and} \quad P(B \cup W) = 1$$

$$P(M) = \frac{40}{80}, \quad P(F) = \frac{40}{80}, \quad \text{and} \quad P(M \cup F) = 1$$

We can calculate the conditional and joint probabilities by using Table 5.7. The conditional probabilities are

$$P(B|M) = \frac{\frac{5}{80}}{\frac{40}{80}} = \frac{5}{40}$$

$$P(W|M) = \frac{\frac{35}{80}}{\frac{40}{80}} = \frac{35}{40}$$

$$P(F|W) = \frac{\frac{15}{80}}{\frac{50}{80}} = \frac{15}{50}$$

$$P(F|B) = \frac{\frac{25}{80}}{\frac{30}{80}} = \frac{25}{30}$$

The joint probabilities are

$$P(B \cap M) = \frac{5}{40}\frac{40}{80} = \frac{5}{80}$$

$$P(B \cap F) = \frac{25}{30}\frac{30}{80} = \frac{25}{80}$$

$$P(W \cap M) = \frac{35}{40}\frac{40}{80} = \frac{35}{80}$$

$$P(W \cap F) = \frac{15}{50}\frac{50}{80} = \frac{15}{80}$$

Suppose we do not know the exact numbers indicated in the table, but we know the joint probabilities and conditional probabilities. We can obtain the marginal probabilities from the joint probabilities by simply summing the joint probabilities.

Table 5.8 Probabilities for stock prices and economic conditions

	Stock price		
Economic condition	*Increase*	*Decrease*	Total
Good	.28	.06	.34.
Normal	.16	.15	.31
Poor	.05	.30	.35
Totals	.49	.51	1.00

For example, if we want to know the probability of selecting a black student, we can sum all the probabilities we know to be associated with black students. The probability of selecting a black student is the probability of selecting a black male student plus the probability of selecting a black female student. That is,

$$P(B) = P(B \cap M) + P(B \cap F) = \frac{5}{80} + \frac{25}{80} = \frac{30}{80}$$

We can calculate other marginal probabilities in the same way:

$$P(W) = P(W \cap M) + P(W \cap F) = \frac{35}{80} + \frac{15}{80} = \frac{5}{8}$$

$$P(M) = P(B \cap M) + P(W \cap M) = \frac{5}{80} + \frac{35}{80} = \frac{1}{2}$$

$$P(F) = P(B \cap F) + P(W \cap F) = \frac{25}{80} + \frac{15}{80} = \frac{1}{2}$$

Hence, the probabilities for individual events—$P(B)$, $P(W)$, $P(M)$, and $P(\text{F})$—are known as marginal probabilities. In this example, say A represents sex and B represents color. The probabilities of individual events can then be represented as $P(A_i)$ and $P(B_i)$ where $i = B, W$ and $j = M, F$.

Example 5.8 Marginal Probabilities on Wall Street. Let A represent the state of economic conditions, and let B represent movement upward or downward of the stock price for Linden, Inc. The probabilities for Linden's stock price movement are presented in Table 5.8.

Let us use I, D, G, N, and P to represent the events of stock price increase, stock price decrease, good economic conditions, normal economic conditions, and poor economic conditions. Then, from Table 5.8, we can calculate the marginal probabilities for the stock price movement of Linden, Inc.:

$$\begin{aligned} P(I) &= P(G \cap I) + P(N \cap I) + P(P \cap I) \\ &= .28 + .16 + .05 \\ &= .49 \end{aligned}$$

$$\begin{aligned} P(D) &= P(G \cap D) + P(N \cap D) + P(P \cap D) \\ &= .06 + .15 + .30 \\ &= .51 \end{aligned}$$

This outcome implies there is a 49 % chance that the stock price will increase and a 51 % chance that the stock price will decrease.

5.6 Independent, Dependent, and Mutually Exclusive Events

Two events are referred to as *independent events* when the probability of one event is not affected by the occurrence of the other. For example, suppose a fair coin is flipped twice. The probability of getting a head on the second toss is not affected by having gotten a head or a tail on the first trial; thus, the two trials are independent. However, many events are not independent. For example, the probability that a child in a less developed country will receive an advanced education is affected by his or her family's economic status. If the family is well-off, the child probably will go on to higher education. Otherwise, the child may have to give up the opportunity for education to help support the family. Therefore, the event of the child's higher education *depends* on the event of his or her family's financial condition.

Suppose a fair coin is tossed once with the probability of 1/2 of obtaining a tail (event A). Let event B be the event of tossing the coin a second time and getting a tail. What is the probability of event B, given event A (one tail)? Or, in symbols, $P(B \mid A)$?

We observe that the occurrence of the second tail is not influenced by (is independent of) the occurrence of the first tail. In such a case, we say event B is statistically independent of event A. *Here*, $P(B|A) = P(B) = .5$, because the occurrence of A has no influence on B.

For independent events, then,

$$P(A \mid B) = P(A) \tag{5.17}$$

$$P(B \mid A) = P(B)$$

$$P(A \cap B) = P(B \cap A) = P(B)P(A) \tag{5.18}$$

Equation 5.18 is a special case of Eq. 5.16. From Eqs. 5.17 and 5.18, we know that the joint probability of two independent events is equal to the product of the marginal probabilities of these two events.

Let A and B be independent such that we know $P(A \cap B) = P(A)P(B)$. Under what circumstances could A and B also be mutually exclusive? If they were, $P(A \cap B)$ would be equal to 0, which implies either $P(A) = 0$ or $P(B) = 0$. Thus, independent events with positive marginal probabilities can never be mutually exclusive.

For example, the GMAT score of student A is independent of the score of student B. The events {A scores ≥ 700} and {B scores ≤ 650} are assumed independent, but they are not mutually exclusive, and two mutually exclusive events cannot be independent.

In sum, a pair of events are mutually exclusive if they cannot jointly occur—that is, if the probability of their intersection is zero.

5.7 Bayes' Theorem

On the basis of Eq. 5.13, we can incorporate additional information into probability analysis. From Eqs. 5.13 and 5.15, we define the conditional probability $P(B|A)$ *as*

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} \tag{5.19}$$

Equation 5.19 represents *Bayes' theorem*, which can be used to incorporate some extra information into the analysis.[3] The most interesting interpretation of Bayes' theorem is in terms of subjective probability, which was discussed in Sect. 5.2.

If we are interested in the event B and form a subjective view of the probability that B will occur, then $P(B)$ is called the *prior probability* in the sense that it is assigned *prior to* the observation of any empirical information. If we later acquire the information that the event A has occurred, this may cause us to modify original judgment about the probability of event B. Because A is known to have occurred, the relevant probability of event B is now the conditional probability of B given A, and it is called the *posterior probability*, or the *revised probability*, because it is assigned *after* the observation of empirical evidence or additional information.

Bayes' theorem provides a method for incorporating new information into our probability beliefs. Formally, we use Bayes' theorem to update a prior probability to a posterior probability when additional information about event A becomes available. We do this by multiplying the prior probability by the adjustment factor $P(A \mid B)/P(A)$.

For example, financial analysts have observed stock prices declining when interest rates increase. They have also observed stock prices moving randomly.

If we collect historical data, we can obtain estimates of the probability of the event "a stock price increase and a decline in interest rates." The probability of a fall in stock prices is what we are most interested in. This probability is called the *prior probability*, because it is based on historical data and our own subjective judgment. If we see the interest rate rise (this is *additional information*), using Bayes' theorem gives us a better estimate of the probability that stock prices will fall. This forecasting method is described in Fig. 5.13.

[3] This theorem is attributed to an English clergyman, the Reverend Thomas Bayes (1702–1761).

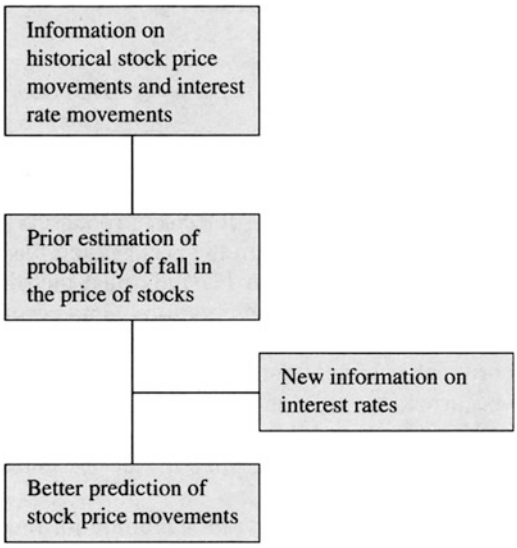


Fig. 5.13 Flow chart of stock price forecasting

Table 5.9 Frequency distribution of changes in stock price

	Interest rate		
Stock price	*Decline*	*Increase*	Unit: frequency
Decline	100	850	950
Increase	900	150	1,050
	1,000	1,000	2,000

To use Eq. 5.19 in the analysis, let

I_D = event that interest rates decline
I_U = event that interest rates increase
S_D = event that stock prices fall
S_U = event that stock prices increase

Then, the conditional probability $P(S_D|I_U)$ can be defined as

$$P(S_D|I_U) = \frac{P(S_D \cap I_U)}{P(I_U)} = \frac{P(S_D)P(I_U|S_D)}{P(I_U)} \tag{5.19a}$$

Table 5.9 summarizes the historical data derived from 2,000 observations of changes in stock price.

From Table 5.9, we can easily estimate $P(S_D|I_U)$ as

$$P(S_D|I_U) = \frac{P(S_D)P(I_U|S_D)}{P(I_U)} = \frac{(950/2,000)(850/950)}{1,000/2,000}$$
$$= \frac{850}{1,000} = .85 \text{ (revised probability)}$$

Using the new information that the interest rate will rise helps us predict a fall in stock prices more accurately. In other words, we are better able to predict the decline in the stock price.

Comparing $P(S_D|I_U)$ with $P(S_D)$ reveals the importance of Bayes' theorem, defined in Eq. 5.19 or Eq. 5.19a. Using Bayes' theorem enables us to make better decisions by incorporating additional information into our probability estimates.

Here, we have discussed using Bayes' theorem for just one basic event. The use of Bayes' theorem for two or more than two events, and the application of this technique in decision making, will be discussed in Chap. 21.

5.8 Business Applications

Application 5.1 Determination of the Commercial Lending Rate. In this example, we show a process for estimating the lending rate a financial institution would extend to a firm (or the lending rate that a borrower would feel is reasonable) on the basis of economic, industry, and firm-specific factors.

In standard banking practice, the lending rate depends in part on the interest rate on government Treasury bills.[4] Therefore, in order to determine the commercial lending rate, we need a forecast of the Treasury bill rate (R_f). This rate will be estimated for three types of economic conditions: boom, normal, and recession.

The second component of the lending rate, (R_p), is the risk premium.[5] It is possible to calculate R_p for each firm by examining the change in *earnings before interest and taxes* (EBIT) under the three types of economic conditions. The EBIT is used as an indicator of the ability of the borrower to repay borrowed funds. Table 5.10 lists all probability information we need to determine the commercial lending rate for Briarworth Company. Column (4) gives the marginal probability, and the probabilities listed in columns (1), (2), and (3) are the joint probabilities.

The probabilities shown in Table 5.10 can also be presented in terms of a tree diagram. Figure 5.14 shows that there are a total of nine possible joint probabilities under the three different economic conditions and the three possible EBIT forecasts.

We can use Table 5.10 and Eq. 5.13 to calculate the conditional probabilities of EBIT level given the economic condition. For example, the probability that Briarworth Company has middle EBIT, given that the economic condition is boom, is

[4] The Treasury bill rate was discussed in Chaps. 2 and 3.

[5] The risk premium is the portion of the interest rate that is above the Treasury bill rate. This additional amount of interest is paid to compensate the lender for the risk it runs in making the loan.

Table 5.10 Probabilities of the lending rate determination for Briarworth Company

	Level of EBIT			
Economic condition	*High (1)*	*Middle (2)*	*Low (3)*	Totals (4)
Boom	.15	.075	.025	.25
Normal	.20	.15	.15	.50
Poor	.025	.05	.175	.25
Totals	.375	.275	.350	1.00

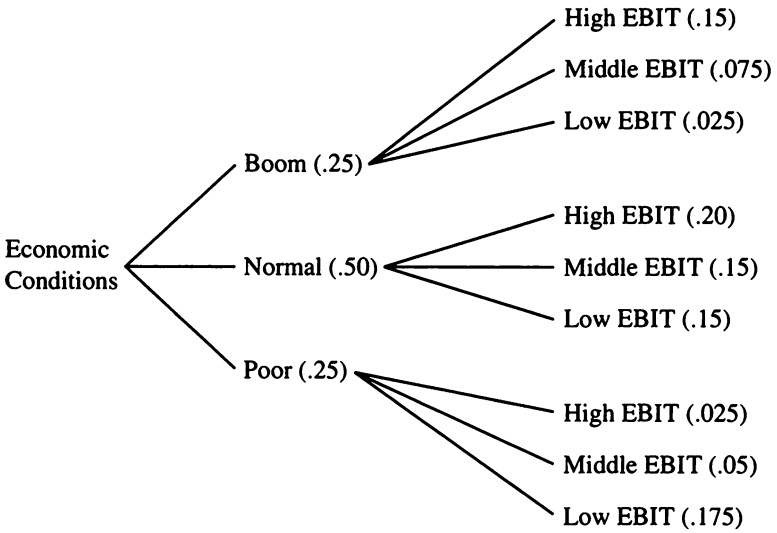


Fig. 5.14 Tree diagram of events for lending rate forecasting

$$P(\text{middle EBIT})|\text{boom}) = \frac{P(\text{boom} \cap \text{middle EBIT})}{P(\text{boom})}$$
$$= \frac{.075}{.25} = .30$$

The conditional probabilities for EBIT level given economic condition are displayed in Table 5.11.

According to Sect. 5.6, a pair of events are independent if and only if their joint probability is the product of their marginal probabilities. In our example, for the events "normal" (normal economic condition) and "middle EBIT," we have, from Table 5.10,

$$P(\text{normal} \cap \text{middle EBIT}) = .15$$

and

$$P(\text{middle EBIT})P(\text{normal}) = (.275)(.50) = .1375$$

Table 5.11 Conditional probabilities of EBIT levels, given economic condition

	EBIT level		
Economic condition	*High*	*Middle*	*Low*
Boom	.60	.30	.10
Normal	.40	.30	.30
Poor	.10	.20	.70

Table 5.12 Worksheet for alternative lending rate estimates for Briarworth Company

Economic Condition	(*A*) R_f	(*B*) Marginal Probability	(*C*) R_p	(*D*) Conditional Probability	(*B* × *D*) Joint Probability	(*A* + *C*) Lending Rate
Boom	12 %	.25	3.0 %	.60	.15	15 %
			5.0	.30	.075	17
			8.0	.10	.025	20
Normal	10 %	.50	3.0 %	.40	.200	13 %
			5.0	.30	.150	15
			8.0	.30	.150	18
Poor	8 %	.25	3.0 %	.10	.025	11 %
			5.0	.20	.05	13
			8.0	.70	.175	16

The product of the marginal probabilities is .1375, which differs from the joint probability .15. Hence, the two events are not statistically independent.

In order to calculate the potential lending rate for Briarworth Company, the loan officer assumes that the predicted Treasury bill rates for boom, normal, and poor economic conditions are 12 %, 10 %, and 8 %, respectively. In addition, the loan officer assumes that the risk premiums for the three different EBIT levels are 3 %, 5 %, and 8 %. Using all this information, the loan officer constructs a worksheet such as Table 5.12.

Table 5.12 shows that during a boom, the Treasury bill rate is assumed to be 12 % but the risk premium can take on different values. There is a 60 % chance that it will be 3.0 %, a 30 % chance it will be 5.0 %, and a 10 % chance it will be 8.0 %. According to the joint probability concepts we discussed in Sect. 5.5, the products of the conditional probability associated with R_p and the marginal probability associated with R_f are the joint probabilities of occurrence for the lending rates computed from these parameters. Therefore, during a boom, there is a 10 % chance that a firm will be faced with a 15 % lending rate, a 7.5 % chance of a 17 % rate, and a 7.5 % chance of a 20 % rate. This process also applies for the other conditions: normal ($R_f = 10$ %) and recession ($R_f = 8$ %).

From Table 5.12, the loan officer for Briarworth Company can estimate the potential lending rates and their probabilities as follows:

Potential lending rate (x_i), %	Probability (P_i), %
20	7.5
18	15.0
17	7.5
16	7.5
15	25.0
13	27.5
11	10.0
	100.0 %

To calculate the estimated average lending rate, we generalized the simple arithmetic average indicated in Eq. 4.2 in Chap. 4 as[6]

$$\bar{x} = \sum_{i=1}^{N} P_i x_i \tag{5.20}$$

where $\sum_{i=1}^{N} P_i = 1$. If P_1 and $P_2 = \ldots = P_N = \frac{1}{N}$, then Eq. 5.20 reduces to Eq. 4.1.

Substituting the information x_i and P_i into Eq. 5.20 yields the estimated average lending rate:

$$\begin{aligned}\bar{x} &= (.20)(.075) + (.18)(.15) + (.17)(.075) + (.16)(.075) \\ &\quad + (.15)(.25) + (.13)(.275) + (.11)(.10) \\ &= 15.1\%\end{aligned}$$

The loan officer can use this estimated average lending rate as a guideline in determining the lending rate. The variance associated with this lending rate and other related analyses will be explored in Example 6.8 and Application 7.4.

Application 5.2 Analysis of a Personnel Data File. The personnel office of the J. C. Francis Company has files for 21,600 employees. These employees are broken down by age and sex in Table 5.13.

If one file is selected at random from the personnel office, what is the probability that it represents:

1. An employee who is 40 years old or younger?
2. A female employee who is 40 years old or younger?
3. Either a male employee or any employee over 40?
4. A male employee over 40?
5. A female employee or any employee 30 years old or older?

[6] We treat x as a measure of central tendency, as discussed in the last chapter. Alternatively, it can be treated as the expected value of a discrete random variable (lending rate), which will be discussed in Sect. 6.3.

Table 5.13 Age and sex classification for J. C. Francis Company

	Sex		
	Female (F)	*Male (M)*	Total
Under 30 (*A*)	3,000	2,500	5,500
30–40(*B*)	4,550	3,800	8,350
Over 40(*C*)	2,850	4,900	7,750
Total	10,400	11,200	21,600

We shall denote the various events involved by A = under 30, B = 30–40, C = over 40, M = male, and F = female:

1. $P(40 \text{ or under}) = P(A \cup B)$

$$= \frac{5,500}{21,600} + \frac{8,350}{21,600} = .6412$$

2. $P(\text{female 40 or under}) = P(A \cap F) + P(B \cap F)$

$$= \frac{3,000 + 4,550}{21,600} = .3495$$

3. $P(\text{male or over 40}) = P(M \cup C) = P(M) + P(C) - P(M \cap C)$

$$= \frac{11,200 + 7,750 - 4,900}{21,600} = .6505$$

4. P (male and over 40) $= P(M \cap C) = \dfrac{4,900}{21,600} = .2269$
5. $P(\text{female or 30 or older}) = P[F \cup (B \cup C)] = P(F) + P(B \cup \text{C}) - P[F \cap (B \cap C)]$

$$= \frac{10,400 + 8,350 + 7,750 - (4,550 + 2,850)}{21,600} = .8843$$

Application 5.3 Soda Purchase Survey. Mr. Mac Francis, manager of a Pathmark Supermarket in central New Jersey, would like to determine:

1. The percentage of Kyle City families that did not purchase any soda during July of 1991
2. The percentage of Kyle City families that purchased either diet or regular soda (or both) during July of 1991
3. The percentage of Kyle City families that purchased only diet soda (no regular soda) during July of 1991
4. The percentage of Kyle City families that purchased only regular soda (no diet soda) during July of 1991
5. Whether diet soda purchases were related to regular soda purchases during the observed month, July of 1991

Table 5.14 Summary table of diet soda and regular soda purchases for Kyle City families during July of 1991

Purchases of six-packs of regular soda (RSODA)	Purchases of six-packs of diet soda (DSODA)		
	0	*1 or more*	Total
0	53	46	99
1 or more	*62*	*39*	*101*
Total	115	85	200

Mr. Francis has asked you to conduct a study to answer these questions. He has provided you with sufficient Kyle City families for you to draw a random sample of 200. You conduct the survey and present Mr. Francis with Table 5.14, which summarizes the data you have accumulated.

The data presented in Table 5.14 make it possible to answer all of Mr. Francis's questions:

1. $P[(\text{DSODA} = 0) \cap (\text{RSODA} = 0)] \dfrac{53}{200} = .265$
   Hence, it is inferred that 26.5 % of Kyle City families did not purchase soda during July of 1991. Note that (DSODA = 0) is an event. It does not imply that P(DSODA) = 0.
2. $P[(\text{DSODA} > 0) \cup (\text{RSODA} > 0)] = P(\text{DSODA} > 0) + P(\text{RSODA} > 0) - \text{P}[(\text{DSODA} > 0) \cap (\text{RSODA} > 0)] = \dfrac{85 + 101 - 39}{200} = .735$
   Consequently, it is inferred that 73.5 % of Kyle City families purchased either diet or regular soda (or both) during July of 1991.
3. $P[(\text{DSODA} > 0) \cap (\text{RSODA} = 0)] = \dfrac{46}{200} = .23$
   Hence, it is inferred that 23 % of Kyle City families purchased only diet soda (no regular soda) during July of 1991.
4. $P[(\text{DSODA} = 0) \cap (\text{RSODA} > 0)] = \dfrac{62}{200} = .31$
   Consequently, it is inferred that 31 % of Kyle City families purchased only regular soda (no diet soda) during July of 1991.
5. Last, we come to the joint probability of (DSODA $\geq$ 1) and (RSODA $\geq$ 1):

$$P[(\text{DSODA} \geq 1) \cap (\text{RSODA} \geq 1)] = \frac{39}{200} = .195$$

$$P(\text{DSODA} \geq 1) \times P(\text{RSODA} \geq 1) = \frac{85}{200} \times \frac{101}{200} = .2146$$

The fact that .195 $\neq$ .2146 implies that the joint probability is not equal to the product of two marginal probabilities. Hence, in accordance with the definition of dependence given in Sect. 5.6, it can be concluded that the purchase of diet

soda was statistically dependent on the purchase of regular soda (and vice versa) during July of 1991.

Application 5.4 Ages and Years of Teaching Experience. Table 5.15 presents the age and number of years of teaching experience of 15 marketing professors. Figure 5.15 is the MINITAB printout of a Venn diagram of set A of marketing professors between 33 and 43 years of age, inclusive, and set B of marketing professors with more than 5 years of teaching experience.

From the Venn diagram, we can calculate the following probabilities:

1. $P(A) = \frac{11}{15} = .73$
2. $P(B) = \frac{7}{15} = .47$
3. $P(A \cap B) = \frac{6}{15} = .40$
4. $P(A \cup B) = \frac{12}{15}$
   $= P(A) + P(B) - P(A \cap B)$
   $= .73 + .47 - .4$
   $= .80$
5. $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{6}{15}}{\frac{7}{15}} = \frac{6}{7} = .857$

Table 5.15 Age and years of teaching experience of 15 marketing professors

Person	Age	Years of teaching experience
1	38	5
2	33	4
3	40	6
4	43	7
5	45	10
6	38	6
7	36	7
8	29	3
9	35	5
10	28	3
11	30	2
12	42	5
13	41	6
14	37	1
15	42	7

```
MTB > NAME C1 'AGE' C2 'EXP'
MTB > READ 'AGE'  'EXP'
DATA> 38 5
DATA> 33 4
DATA> 40 6
DATA> 43 7
DATA> 45 10
DATA> 38 6
DATA> 36 7
DATA> 29 3
DATA> 35 5
DATA> 28 3
DATA> 30 2
DATA> 42 5
DATA> 41 6
DATA> 37 1
DATA> 42 7
DATA> END
15 rows read.
MTB > PRINT 'AGE' 'EXP'
```

Data Display

```
Row    AGE    EXP
 1      38      5
 2      33      4
 3      40      6
 4      43      7
 5      45     10
 6      38      6
 7      36      7
 8      29      3
 9      35      5
10      28      3
11      30      2
12      42      5
13      41      6
14      37      1
15      42      7
MTB > GSTD

* NOTE  *      Standard Graphics are enabled.
               Professional Graphics are disabled.
               Use the GPRO command to enable Professional Graphics.

MTB > PLOT 'EXP'  'AGE'
```

Character Plot

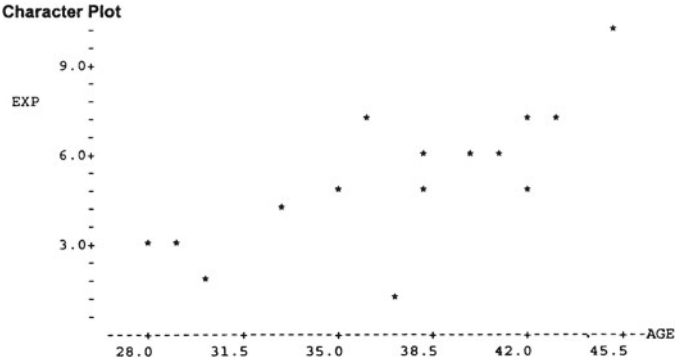


Fig. 5.15 MINITAB printout of Venn diagram for age and years of teaching experience

5.9 Summary

In this chapter, we developed some of the basic tools of probability. The concept of probability enables us to assess the probabilities of various sample outcomes, given a specific population structure. In addition to discussing the basic concepts of probability, we explored more advanced topics such as conditional probability, joint probability, and marginal probability. We also showed how it is possible to use additional information to update probabilities by applying Bayes' theorem.

In Chap. 6, we extend the topics discussed in this chapter by introducing the concepts of discrete random variables and probability distributions. And in Chaps. 7 and 9, we extend these concepts to the case of continuous random variables.

Questions and Problems

1. Two cards are drawn from an ordinary deck of shuffled cards.
   (a) What is the probability that they are both queens if the first card is replaced?
   (b) What is the probability that they are both queens if the first card is not replaced?
2. Find the probability of a 5 turning up at least once in 2 tosses of a fair die.
3. Find the probability of rolling a 1 on the first roll of a die, a 2 on the second, and a 3 on the third.
4. A bag consists of ten balls, three white and seven red.
   (a) What is the probability of drawing a white ball?
   (b) What is the probability of drawing a white ball on the first draw and a red ball on the second draw when the first ball is replaced?
   (c) How would your answer to part (b) change if there were no replacement?
5. You are given the sample space $S = \{a, b, c, d, e\}$ and the events $A = \{a, c, e\}$ and $B = \{b, d, e\}$.
   (a) List the events $A \cup B$, $A \cap B$, $\bar{A}$, $\bar{B}$, $\bar{A} \cap B$, and $(\overline{A \cup B})$.
   (b) Draw a Venn diagram for $A \cap B$.
6. Suppose you are flipping a fair coin.
   (a) Find the probability of flipping four heads in a row.
   (b) Find the probability of flipping H T H T H.
   (c) Find the probability of flipping five heads in six flips.
7. What is the probability that at least two students in a class of 20 students will have the same birthday? Assume that there are no twins among the students and that all of the 365 birthdays are equally likely.
8. What is the probability of drawing three spades in a row from a standard deck of cards without replacement?

9. Roll a pair of dice ten times and then calculate the mean and standard deviation for these rolls.

   (a) What is the largest possible mean?
   (b) What is the smallest possible mean?
   (c) What is the smallest possible standard deviation?

10. Suppose a bag contains 12 balls distributed as follows: five red dotted balls, two red striped balls, one gray dotted ball, and four gray striped balls.

    (a) Suppose you draw a red ball from the bag. What is the probability that it is striped?
    (b) Suppose you draw a gray ball. What is the probability that it is dotted?
    (c) Suppose you draw a dotted ball. What is the probability that it is red?

11. Determine the probability of betting on a winning number in a game of roulette. The numbers on the wheel are 0, 00, and 1 through 36. Each number is as likely as any other to become a winning number.
12. The probability that a car dealer will make a sale when he meets a prospective customer is 20 %. If he meets three customers at random, what is the probability that all three customers will purchase a car?
13. Find the following joint probability: the probability that a sale will result in a sales commission, given that 75 % of the sales representatives receive a commission on their sales and that 80 % of the company's sales are made by sales reps.
14. Roll a die 25 times and construct a table showing the relative frequency of each of the 6 possible numbers.
15. Calculate the probability for scores less than 600.

SAT scores for Fiesta University

SAT	Number of students
750–800	40
700–750	60
650–700	100
600–650	250
550–600	375
500–550	575
450–500	400
<450	100

16. In which of the following sets are the two events independent? In which are they mutually exclusive? In which are they neither?

    (a) The Detroit Pistons win the NBA champion ship and the Oakland A's win the World Series.
    (b) The Boston Red Sox win the pennant and the Boston Red Sox sell more than two million tickets in the same season.
    (c) Both the New York Mets and the New York Yankees win the World Series in 1995.

(d) Both the New York Mets and the New York Yankees win the pennant in 1995.

17. A bag contains three balls: a black one, a white one, and a red one. A magician takes the balls out one by one. Draw an outcome tree. What is the probability of drawing the balls in the order of red-white-black?
18. Suppose that $P(E_1) = 0.3$ and $P(E_2) = 0.4$. Obtain $P(E_1 \cup E_2)$, $P(E_1|E_2)$, and $P(E_2|E_1)$, given that $P(E_1 \cap E_2) = .1$.
19. A baseball player has a lifetime batting average of .3. During a game, he has 5 at bats. A student of statistics argues that his chance of going 5 for 5 is $(.3)^5$. Do you agree? What assumption does the student make to come up with this answer?
20. The sales department wants to send two sales representatives on a business trip. There are five women and five men in the department. If the sales manager randomly selects two people, what is the probability that one woman and one man will be picked?
21. A survey of 200 students yields the following data:

	Own TV	Do not own TV
Own computer	50	30
Do not own computer	80	40

(a) What is the probability of drawing at random a student who owns both a computer and a TV?
(b) What is the probability of drawing at random a student who owns only a computer?
(c) In part (b), suppose we draw a student who owns a computer. What is the probability that this student also owns a TV?
(d) What is the marginal probability of owning a computer?

22. In question 21, if we draw two students at random without replacement, what is the probability of our getting a student who owns both and a student who owns neither?
23. A cereal company runs a certain advertisement in three media: newspaper, radio, and TV. Of the customers surveyed, 40 saw the advertisement on TV, 40 heard it on the radio, 30 read it in the newspaper, 20 saw it both in the newspaper and on TV, 20 know it both from TV and radio, 20 know it both from newspaper and radio, and 10 know it from all three media. How many customers were surveyed? (*Hint:* Use a Venn diagram.)
24. A city company has 300 employees. Among these employees, two out of every three take public transportation to work, one out of every two owns a car, and one out of every three owns a car but takes public transportation to work. How many employees do not own a car and take public transportation to work?
25. Draw 2 cards from a deck of cards without replacement. What is the probability of getting a diamond on the first draw and a club on the second? What is the

probability of drawing 2 cards and getting a diamond and a club regardless of order?

26. What is the probability of getting the same outcome in 2 rolls of a die? What is the probability that the sum of 2 outcomes is 7?
27. Of the light bulbs delivered on May 25, 400 are produced in the morning shift, 300 in the evening shift, and 300 in the night shift. Say we pick a light bulb at random.

    (a) What is the probability that we have a light bulb produced in the night shift?
    (b) What is the probability that we have a light bulb produced in either the morning shift or the night shift?

28. In question 27, assume that $\frac{1}{10}$ of the light bulbs produced in the morning shift, $\frac{1}{10}$ of the light bulbs produced in the evening shift, and $\frac{1}{5}$ of the light bulbs produced in the night shift are detective. Say we pick a light bulb at random.

    (a) What is the probability that the light bulb is defective?
    (b) What is the chance that the light bulb is defective and was produced in the night shift?
    (c) Suppose we get a defective light bulb. What is the chance that this light bulb was produced in the night shift?

29. A sports magazine wants to learn something about its subscribers. The subscribers are classified as teenagers or older people and as being in school or holding a job. The magazine sends out a questionnaire to its readers and obtains the following results:

    40 % are older than 20.
    60 % are teenagers.
    40 % of the teenagers who subscribe are in school.
    40 % of the subscribers hold a job.

    What is the possibility that a subscriber is older than 20 and holds a job?

30. The business majors at Metropolitan University can be broken down as follows:

Major	Male	Female	Total
Accounting	200	400	600
Finance	400	250	650
Marketing	200	250	450
Management	200	100	300
Total	1,000	1,000	2,000

    (a) We have randomly selected four students to attend a regional conference. What is the probability that we have a representative from each major?
    (b) We have randomly selected four students to attend a regional conference. What is the probability that we have a female student from each department?

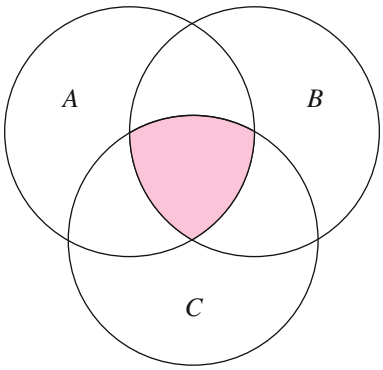
(c) The dean has randomly selected a student from each department to attend a regional conference. What is the probability that all four students selected are females?

31. A survey at Metropolitan College shows that among 750 economics majors, every student has taken at least one course in either economics or statistics. We also know that:

    450 students have taken statistics.
    450 have taken microeconomics.
    450 have taken macroeconomics.
    250 have taken both micro- and macroeconomics.
    200 have taken both microeconomics and statistics.
    250 have taken both macroeconomics and statistics.

    (a) How many students have taken all three courses?
    (b) What is the probability that a student who we know has taken a course in microeconomics has also taken statistics?

32. A local factory has two shifts: day shift and night shift. The day shift produces $\frac{2}{3}$ of the total product. Of the day shift product, 1 % are defective. Of the night shift product, 2 % are defective. If we randomly select one product, what is the chance that it was produced during the day shift? If the selected product is defective, what is the probability that it was produced during the night shift?
33. The following picture helps you obtain the probability that events A, B, and C, happen jointly. Write down the formula for obtaining $P(A \cap B \cap C)$.



34. A hospital found that the probability of a power failure in a certain time period is .00001. To guarantee the functioning of the hospital, the hospital installed a backup system that has a probability of .005 of breaking down. The two power systems operate independently. What is the probability that the hospital will completely stop functioning?
35. An insurance agent talks to three customers each day. Her probability of making a sale in the first meeting with a customer is .2. When she gets the second meeting with the same customer, the probability of making a sale

increases to .8. In the past 2 days, this agent has talked to three customers twice. What is the probability that she made no sales?

36. Three different manuals were used to teach students how to type. Each manual was used by $\frac{1}{3}$ of the students. The results show that 30 % of the students using manual A, 20 % of the students using manual B, and 10 % of the students using manual C can pass a typing test. We have found a student who passed the test. What is the probability that this student used manual A?
37. Fifty percent of the economists in the country are conservatives. The other 50 % are liberals. Thirty percent of the conservative economists and 20 % of the liberal economists believe that we will have a recession. We have found an economist who thinks we will see a recession in the next year.
What is the probability that he or she is a conservative economist?
38. A training program is effective for 80 % of the students whose mathematics background is strong, but it is effective for only 60 % whose math background is not good. Assume that only 60 % of a group of students are well trained in mathematics. What is the chance that the training program will be effective for this group of students?
39. A training program is effective for 80 % of the students who are strongly motivated, but it is effective on only 60 % of the students who are not strongly motivated. Assume that only 60 % of the students are strongly motivated. We have selected a student who has benefited from the program. What is the probability that this student was strongly motivated?
40. Three percent of the products produced by the new assembly line are defective. Five percent of the products produced by the old machine are defective. The new machine produced 70 % of the total product. The old machine produced 30 % of the total product. We randomly draw a product and discover that it is defective. What is the probability that this defective item was produced by the old machine?
41. Mr. Doe wants to send two employees in his company on a business trip. He has five employees in the company. In how many different ways can he organize the trip?
42. When playing the Megabucks Lottery, a player is supposed to pick 6 numbers out of 48. If the lottery committee randomly picks the same 6 numbers, then the player hits the jackpot. What is the chance that a player will hit the jackpot?
43. An advertising agency wanted to find out what kinds of readers subscribed to a sports magazine. The agency sent out questionnaires with the magazine and received the following result:

	Blue-collar job	White-collar job
Teenagers	20	30
The middle aged	30	30
Old people	30	10

(a) If we know that a reader is a blue-collar worker, what is the probability that this reader is also an old person?

(b) If we know that a reader is a teenager, what is the probability that this reader is also a white-collar worker?

44. Define the following terms: event, random experiment, subset, sample space, sample points.
45. Why is the concept of probability important to understanding statistics?
46. Explain what we mean when we say two events are independent.
47. Compare a simple event to a composite event. Give an example of each.
48. What do we mean by the union of two events? What do we mean by the intersection of two events? Use a Venn diagram to illustrate this point.
49. Explain what we mean by mutually exclusive events.
50. Briefly define conditional probability. Give some examples of conditional probability.
51. What is a joint probability? What is a marginal probability?
52. What is a prior probability? What is a posterior probability? Briefly explain how Bayes' theorem can be used to link the two.
53. What is the probability of obtaining a head in one toss of a fair coin? What is the probability of rolling a 5 in one roll of a fair die? What is the probability of tossing a head and rolling a 5?
54. You are dealt 4 cards from a standard 52-card deck. What is the probability that you will be dealt all 4 aces?
55. Again, you are dealt 4 cards. The first card is a spade, the second a heart, the third a diamond, and the fourth a club. What is the probability that you are dealt all 4 aces?
56. Consider the roll of a 6-sided die, its faces numbered 1, 2, 3, 4, 5, and 6. Draw a Venn diagram showing the six possible outcomes. Now draw circles showing the following rolls:

    (a) An odd number
    (b) 2 or an odd number
    (c) 3 or an even number

57. Again, consider the roll of a 6-sided die. Given the following events A and B, find the intersection and the union for A and B if

    (a) $A = \{1,3, 5\}$ and $B = \{2,4,6\}$
    (b) $A = \{1,3\}$ and $B = \{1,3,5\}$
    (c) $A = \{1,2, 3\}$ and $B = \{2,4,5\}$
    (d) $A = \{1,2, 3, 4\}$ and $B = \{3,4,5,6\}$

58. You have drawn three diamonds, one spade, and one heart from a deck of cards. If you discard the spade and the heart, what is the probability of your drawing two cards from the remaining 47 cards to obtain a flush (five cards of the same suit)?
59. In poker, a royal flush consists of A, K, Q, J, and ten of the same suit. What is the probability of drawing five cards and obtaining a royal flush? What is the probability of being dealt a royal flush in spades?

60. Suppose there are six unrelated people in a room. What is the probability that any two of them have the same birthday?
61. Suppose you toss a coin three times. What is the probability of tossing three heads in a row? What is the probability of tossing three tails in a row? What is the probability of tossing either three heads or three tails in a row?
62. An advertising executive decides that a television commercial should be shown on two television stations. If three television stations serve the area the company wants to reach, how many possible combinations does the executive have to choose from? If a fourth television station becomes available, how many combinations are there now?
63. An automobile manufacturer produces cars in four different colors and offers three different options packages. How many different combinations of color and options package can the auto manufacturer offer?
64. A basketball player makes 75 % of his shots from the foul line. What is the probability of this player making 10 shots in a row from the foul line? Are there any assumptions we need to make to compute this answer?
65. Your investment advisor has a portfolio of 75 stocks: 40 high-growth stocks and 35 high-dividend stocks. Of the 40 high-growth stocks, 25 have increased in value over the last year, whereas 10 of the high-dividend stocks have increased in value.

    (a) If a stock is selected at random, what is the probability that the stock will be a high-dividend stock that has increased in value?
    (b) What is the probability of selecting a stock that has not increased in value?
    (c) If the stock selected has increased in value, what is the probability that it is a high-growth stock?

66. The Whiter Smile Company is about to begin selling a new toothpaste. Company planners know that the probability of the new product being profitable is 10 %. They also know from previous market research that when their test panel likes the product, it has an 80 % chance of being profitable. Historically, panels like 10 % of the new products. Using the Bayesian approach, find the probability that the panel liked the product if the toothpaste is profitable.
67. Suppose you flip a fair coin once and roll a 6-sided die once. What is the probability of tossing a tail and rolling a one?
68. Suppose you flip a coin twice and roil a die twice. What is the probability that you will toss 1 head and 1 tail and will roll two 6's?
69. A top amateur bowler has a 70 % chance of roiling a strike. What is the probability that this bowler will bowl a perfect game (12 strikes in a row)? Are there any assumptions we need to make to answer this question?
70. The Tastee Coffee Company is about to begin selling a new gourmet coffee. Company managers know that the probability the product will be profitable is 20 %. They also know that the probability that the test panel will like the product is 20 %. They also know from previous market research that when the product is profitable, there is a 60 % chance that their test panel liked the product and that when it is unprofitable, there is a 95 % chance that the panel

did not like it. Using the Bayesian approach, find the probability that the gourmet coffee is profitable if the test panel liked the product.

71. A real estate developer offers homes in five different colors and three different models. How many different combinations of color and model can the real estate developer offer?
72. Rah Rah College has a limited number of dorm rooms available to students, so every year students participate in a lottery to determine whether they will have school housing next year. Suppose that every year 25 % of the students do not receive school housing. In his sophomore year, Bob Smith is one of the "losers" in the lottery and does not receive school housing. He consoles himself by noting that because 25 % of the students do not receive housing each year "everyone should lose once in the 4 years of college." He therefore figures that he is assured of getting housing in his junior and senior years. Is Bob's assumption accurate?
73. A stock broker owns five suits and 12 ties. Assuming that all the suits and ties match, determine how many different outfits (combinations) the stock broker can wear.
74. An advertising agency suggests that a bicycle manufacturer advertises in four of the seven bicycling magazines. How many different combinations of four magazines can be selected?
75. A Senate committee consists of six Democrats and five Republicans. In how many ways can a subcommittee consisting of four Democrats and four Republicans be formed?
76. You believe you have come up with a fool-proof way to win at roulette. Because the odds are nearly 50–50 that red will come up and nearly 50–50 that black will come up in roulette, you believe that whenever black comes up, red will come up next, and vice versa. Do you think this is a winning strategy?
77. At the beginning of each week, a company decides how much to spend on newspaper ads for that week. It spends either \$250 or \$500 each week on newspaper ads. Assuming there is an equal probability of spending either amount, find the probability that in a month (4 weeks), the total advertising expenditure is greater than \$1,250.
78. A car salesman meets 12 customers each week. His probability of making a sale in the first meeting with a customer is .3. When he gets the second meeting with the same customer, the probability of making a sale increases to .9. Over the last 2 weeks, he talked to four customers twice. What is the probability that he made no sales?
79. A life insurance company knows with certainty that all people will die someday. Does it make sense for the insurance company to use probability theory to set its life insurance rates?
80. Consider the sample space $S = \{A, B, C, D, E, F, G\}$ and the following events:

$I = \{A, C, E, G\}$
$II = \{B, D, E\}$
$III = \{A, B, C, D\}$

$IV = \{E, F, G\}$
$V = \{B, F\}$
$VI = \{B, D, E, F\}$

Are the following sets of events mutually exclusive, collectively exhaustive, both, or neither?

(a) *I* and *II*
(b) *III* and *IV*
(c) *I* and *III*
(d) *II* and *IV*
(e) *I* and *IV*
(f) *II* and *III*
(g) *I* and *V*
(h) *I* and *VI*

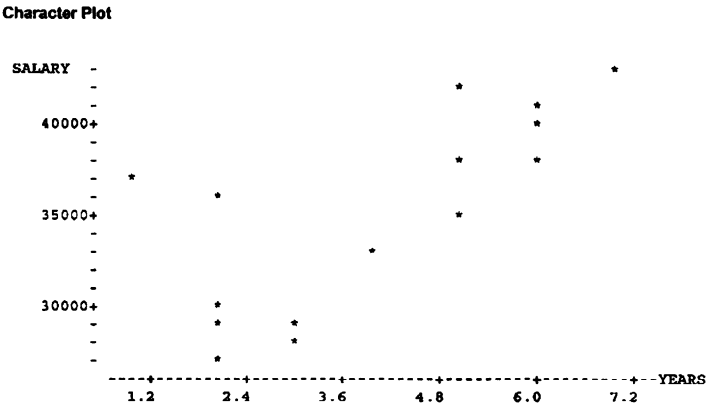
81. State the complement of each of the following events:

(a) Drawing a spade from a full deck of cards
(b) Inflation of less than 5 % per year
(c) GNP growth of more than 4 % per year

82. The figure below is a plot of salary and experience of employees of the Endicott Company. Answer the following questions by using a Venn diagram.

(a) The probability of experience between 3 and 5 years
(b) The probability of more than 4 years' experience and a salary of between $25,000 and $37,000

```
MTB >  NAME C1 'YEARS' C2 'SALARY
MTB > READ 'YEARS'  'SALARY'
DATA> 5 38000
DATA> 4 33000
DATA> 6 40000
DATA> 7 43000
DATA> 2 36000
DATA> 6 38000
DATA> 2 29000
DATA> 3 29000
DATA> 5 35000
DATA> 3 28000
DATA> 2 30000
DATA> 8 42000
DATA> 6 41000
DATA> 1 37000
DATA> 2 27000
DATA> END
      15 rows read.
MTB > GSTD
* NOTE  * Standard Graphics are enabled.
           Professional Graphics are disabled.
           Use the GPRO command to enable
           Professional Graphics.
MTB > PLOT 'SALARY'  'YEARS'
```



83. All the students of a university are assigned ID numbers. The ID number consists of the first three letters of a student's last name, followed by four numbers. How many possible different ID numbers are there?
84. Suppose there are four events A, B, C, and D. The following information is given:

$P(A) = .5$	$P(A \bigcup D) = .72$
$P(B) = .15$	$P(A \mid B) = .25$
$P(C) = .20$	$P(A \bigcap C) = .04$
	$P(A \bigcap D) = .03$

(a) Compute P(D).
(b) Compute $P(A \mid D)$.
(c) Compute $P(A \bigcap B)$.
(d) Compute $P(A \bigcup B)$.
(e) Are A and B mutually exclusive? Explain your answer.
(f) Are A and B independent? Explain your answer.

85. Assume you have applied to two different universities A and B. In the past, 30 % of students who applied to University A were accepted, while University B accepted 45 % of the applicants. Assume events are independent of each other.

(a) What is the probability that you will be accepted in both universities?
(b) What is the probability that you will be accepted to at least one graduate program?
(c) What is the probability that one and only one of the universities will accept you?
(d) What is the probability that neither university will accept you?

86. Suppose 20 % of the employees of company ABC have only a high school diploma, 60 % have bachelor degrees, and 20 % have graduate degrees.

Of those with only a high school diploma, 15 % hold management positions; whereas, of those having bachelor degrees, 30 % hold management positions. Finally, 60 % of the employees who have graduate degrees hold management positions.

(a) What percentage of employees holds management positions?
(b) Given that a person holds a management position, what is the probability that she/he has a graduate degree?

Appendix 1: Permutations and Combinations

In some probability problems, we encounter a finite set with n distinct elements (objects) $\{e_i, i = 1, 2,\ldots, n\}$ in the sample space:

$$S = \{e_1, e_2, \ldots, e_n\} \tag{5.21}$$

If we want to know how many different ways there are of ordering these elements, then using permutation and combination techniques is the most effective way to proceed. For example, we know that 10 % of Wakeley Company's accounts receivable contain errors. If 6 are selected at random, with replacement, then we can use permutation and combination techniques to calculate the probability that exactly two of those selected contained errors. (The solution of this problem appears in Example 5.10.)

Permutations

The number of distinct arrangements that can be made from n elements of S, using r of them at a time, is denoted by ${}_nP_r$ and is called the *number of permutations* of n things taken r at a time ($r \leq n$). The number of permutations of a set of objects represents the number of ways the objects can be ordered. To obtain the result of ${}_nP_r$, we can apply the basic counting rule to the coin-tossing case. If a coin was tossed four times, then there are four steps (tosses), and each toss has two possible outcomes (heads and tails). The total number of outcomes in the experiment (N) is $N = 2 \cdot 2 \cdot 2 \cdot 2 = 16$.

To generalize this type of calculation, suppose we denote the number of outcomes in the first step of the experiment as n_1, the number of outcomes in the second step as n_2, and so on, where n_k denotes the number of outcomes in the last (the kth) step. The basic counting rule states that the total number of outcomes (N) equals the product of the numbers of outcomes in all steps:

$$N = n_1 \cdot n_2 \cdot n_3 \cdot \cdots \cdot n_k \tag{5.22}$$

Suppose we have some number r of objects that are to be placed in order, and suppose each object can be used only once. How many different sequences are possible? This problem is similar to that defined in Eq. 5.22. It can readily be shown

that $n_1 = r$, $n_2 = r - 1$, …, $n_r = 1$. Hence, the number of possible orderings of r objects is

$$r! = (r)(r-1)\cdots(2)(1) \tag{5.23}$$

Equation 5.23 represents a factorial product.

Suppose now that we have n objects from which r are to be selected. The number of ways in which it is possible to select the r objects can be determined from the following product:

$$n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot (n-r+1)$$

where

n = number of choice for the first object
$n - 1$ = number of choice for the second object
$n - 2$ = number of choice for the third object
$n - r + 1$ = number of choice for the $(n - r + 1)$th object

Thus, the permutations ${}_nP_r$ can be defined as

$$\begin{aligned} {}_nP_r &= n(n-1)\cdots(n-r+1) \\ &= \frac{n!}{(n-r)!} \end{aligned} \tag{5.24}$$

For example, say we want to know in how many arrangements we can assign four students to three seats. We can put any of the four students in the first seat; there are four possibilities here. Then, we can put any of the remaining three in the second seat. Finally, we must choose between the remaining two for the third seat. Thus, ${}_4P_3 = (4)(3)(2) = 24$. This example illustrates that the "order," or arrangement, is important for a permutation.

Example 5.9 Permutations of the Letters A, B, and C. We are given the three letters A, B, and C. To determine the number of possible arrangements, note that we have three ways to select the first letter. Once the first letter has been selected, there are two ways to select the second letter from those that remain. There is only one way to select the third letter. Of course, no letter can be selected more than once in any arrangement. Using Eq. 5.22, we find that the total number of ways to make the selection (i.e., to arrange the letters in order) is

$$3! = 3 \cdot 2 \cdot 1 = 6$$

Figure 5.16 is a tree diagram showing six possible arrangements of the three letters. Each arrangement is a branch of the tree.

Combinations

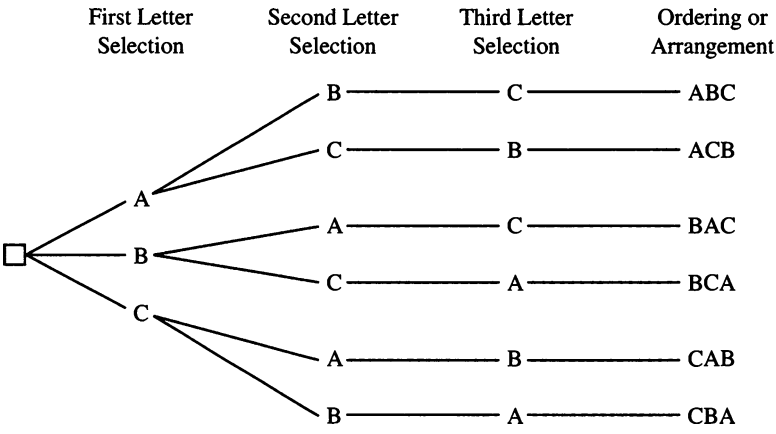


Fig. 5.16 Tree diagram: permutations of the letters A, B, and C

The number of permutations of a set of objects represents the number of ways the object can be ordered. Suppose we are interested in the number of different ways in which r objects can be selected from n, but we are not concerned with the order. Then, the number of possible selections is called the *number of combinations* and is denoted by $\binom{n}{r}$. It can be shown that $\binom{n}{r}$ and ${}_nP_r$ are related by formula

$$r!\binom{n}{r} = {}_nP_r = \frac{n!}{(n-r)!}$$

Therefore,

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{5.25}$$

For example, if $n = 5$ and $r = 3$, then

$$\binom{n}{r} = \frac{5!}{3!(5-3)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1)(2 \cdot 1)} = 10$$

Permutations and combinations can be used to simplify probability expressions and facilitate their evaluation.

Example 5.10 Errors of Accounts Receivable. Suppose 10 % of Wakeley Company's accounts receivable are known to contain errors. If six accounts receivable are selected at random, with replacement, what is the probability that:

1. None of those selected contains an error?
2. Exactly two of those selected contain errors?
3. At most two of those selected contain errors?
4. At least two of those selected contain errors?

Solutions

1. $P(\text{no errors}) = (9/10)^6 \approx .531$
2. We consider first the probability that the first two accounts receivable chosen contain errors and the remaining four do not. This is given by

$$\frac{1}{10}\cdot\frac{1}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10}\cdot\frac{9}{10} = \left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^4$$

But there are $\frac{6}{2} = \frac{6}{4}$ combinations of the accounts with errors and four without. Each of these arrangements occurs with the probability

$$\left(\frac{1}{10}\right)^2 \cdot \left(\frac{9}{10}\right)^4$$

Thus, by using Eq. 5.25, we obtain

$$\begin{aligned} P(\text{exactly two errors}) &= \binom{6}{2}(1/10)^2(9/10)^4 \\ &= \frac{6!}{2!4!}\cdot(.1)^2(.9)^4 \\ &= (15)(.0066) \\ &= .098 \end{aligned}$$

3. By repeatedly employing the binomial formula as discussed in Part 2, we obtain

$$\begin{aligned} P(\text{at most } 2) &= P(\text{exactly } 2) + P(\text{exactly } 1) + P(0) \\ &= \binom{6}{2}(1/10)(9/10)^4 + \binom{6}{1}(1/10)(1/9)^5 + \binom{6}{0}(9/10)^6 \\ &= .098 + .354 + .531 \\ &= .984 \end{aligned}$$

4. $P(\text{at least } 2) = P(\text{exactly2}) + P(\text{more than } 2)$

$= .098 + [1 - P(\text{at most } 2)]$
$= .098 + 1.000 - .984 = .114$

Of course, this problem can also be solved by direct computation similar to the method used in Part 3.

Example 5.11 The Birthday Problem. To compute probabilities, we often need to understand the concept of permutations and combinations. The "birthday problem" is a popular example of probability based on permutations. In the birthday problem, we are interested in the probability that at least two people in a given room have the

same birthday. As we increase the number of people in the room, the number of possible combinations of people increases. With only two people in a room, there is only one possibility for a match. With three people (A, B, and C) in a room, there are three possible matches: A with B, A with C, and B with C. With four people in a room there are six possible matches and so on.

What is the probability that in a class of $m = 50$ students, at least two students have the same birthday? To solve this, we assume that there are not twins among the m people in the class and that each of the 365 possible birthdays is equally likely. We therefore assume that anyone born on February 29 (leap year) will consider her or his birthday to be March 1 to make the problem manageable.

On the basis of these assumptions, we can see that there are 365 possible birthdays for each of the m people. Therefore, the sample space contains 365^m outcomes, all of which are equally probable. Now we proceed as though we were interested in the probability that *no* two people have the same birthday. We divide the number of permutations by the total number of outcomes. Letting B represent the event of m students having different birthdays is precisely the same as asking in how many ways m birthdays can be selected from 365 possible birthdays and arranged in order. This is just the number of permutations, ${}_{365}P_m$. Then,

$$P(\bar{B}) = \frac{{}_{365}P_m}{365^m}$$

is the probability that of our m people, no two have the same birthday. This is the complement of event B, at least two people having the same birthday.

Using Eqs. 5.11 and 5.24, we find that the probability P that out of m students at least two people have the same birthday is

$$P(B) = 1 - \frac{{}_{365}P_m}{365^m} = 1 - \frac{365!}{(365-m)!365^m} \tag{5.26}$$

The following table shows the probability (P) for different values of m. We can see that with only 50 students in a class, there is a 97 % probability that two or more students will have the same birthday.

m	P
10	.117
20	.411
30	.706
40	.891
50	.970
100	.9999997

Outcome Trees and Probabilities

The probability of an outcome is often much more difficult to calculate than that of the outcome of rolling a die once. For example, consider a biased coin that has a 1/3 probability of coming up heads and a 2/3 probability of coming up tails. If we

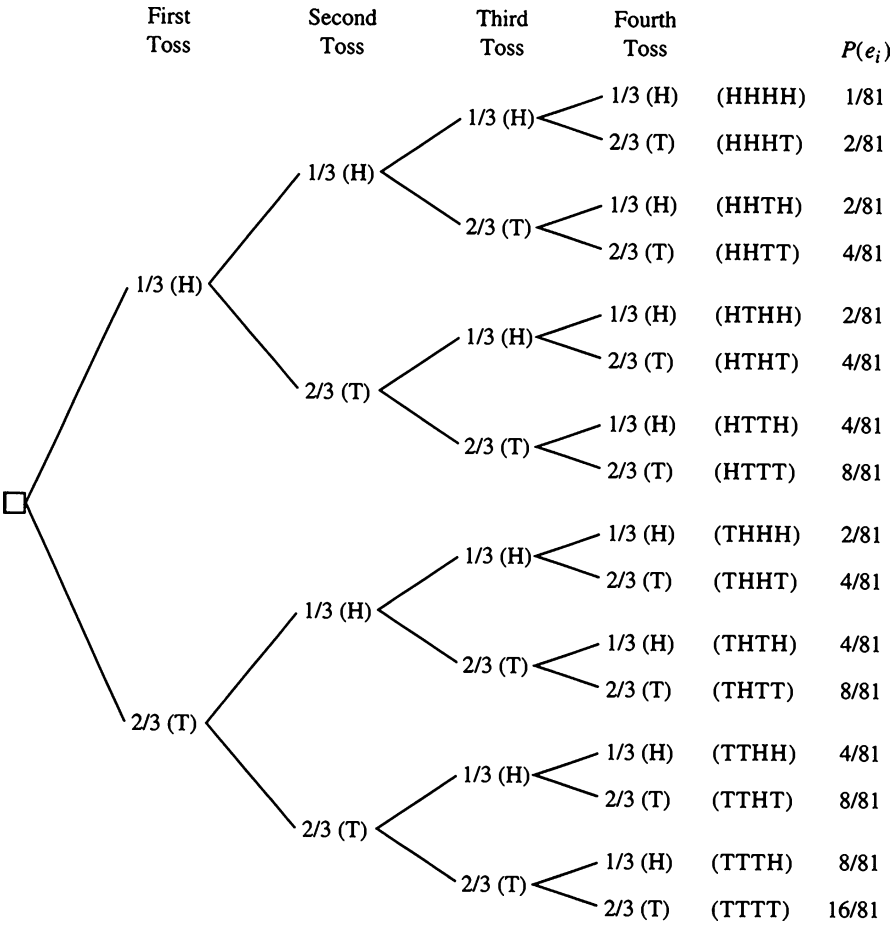


Fig. 5.17 An outcome tree for four tosses of a biased coin

flip this biased coin four times and list the possible outcomes toss by toss, we obtain the results shown in Fig. 5.17, which make up an *outcome tree*.

Let's consider the possible outcomes listed in the fifth column of Fig. 5.17. There are 16 distinct possible outcomes, which can be represented as

$$\{e_1, e_2, e_3, \ldots, e_{16}\}$$

where e_1 = (HHHH), e_2 = (HHHT), …, e_{15} = (TTTH), and e_{16} = (TTTT). Using the relative frequency concept of probability as indicated in Eq. 5.2 in the text, how do we find the probability of, for example, e_1 = (HHHH)? If the probability of an individual outcome is independent, we can find $P(e_1)$ by multiplying together the probabilities of all outcomes. An event is said to be independent if its outcome does not depend on past outcomes in this case. For example, from our coin-tossing

experiment, the probability of tossing a head is always 1/3, regardless of the previous toss. So the probability of tossing two heads in a row is the probability of tossing a head on the first toss multiplied by the probability of tossing a head on the second toss. Thus, the probability of getting four heads is

$$\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{1}{3}=\frac{1}{81}$$

and the probability of getting two heads first and two tails later is

$$\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{2}{3}\cdot\frac{2}{3}=\frac{4}{81}$$

From Fig. 5.17, we can calculate the probability of the event consisting of three heads and one tail as

$$\frac{2}{81}+\frac{2}{81}+\frac{2}{81}+\frac{2}{81}=\frac{8}{81}$$

Alternatively, this probability can be calculated as follows:

$$\binom{n}{r}(p)^{r}(1-p)^{n-r}=\binom{4}{3}\left(\frac{1}{3}\right)^{3}\left(\frac{2}{3}\right)=\frac{4!}{3!1!}\left(\frac{2}{81}\right)=\frac{8}{81} \tag{5.27}$$

Equation 5.27 represents a binomial combination formula for calculating the probability. This formula was discussed in Example 5.10. The concept will be used in developing binomial distribution in the next chapter.

Chapter 6
Discrete Random Variables and Probability Distributions

Chapter Outline

Key Terms

Random variable	Hypergeometric distribution
Discrete random variable	Hypergeometric random variable
Continuous random variable	Hypergeometric formula
Probability function	Poisson distribution
Probability distribution	Joint probability function
Probability mass function	Joint probability distribution
Cumulative distribution function	Marginal probability function
Step function	Conditional probability function
Expected value	Conditional probability distribution
Bernoulli process	Covariance
Binomial distribution	Coefficient of correlation
Binomial probability function	Option
Lot acceptance sampling	Random walk

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_6,

6.1 Introduction

In Chaps. 2, 3, and 4, we explored descriptive statistical measures, and we examined probability concepts and techniques in Chap. 5. Here we will build on this foundation as we establish the definitions of discrete and continuous random variables and discuss important discrete probability distributions in terms of specific numerical outcomes.

The binomial distribution, hypergeometric distribution, Poisson distribution, and joint probability functions are discussed in detail in this chapter. We also explore the Poisson approximation to the binomial distribution and examine joint probability functions and distributions. Finally, we investigate expected value and variance of the sum of both uncorrelated and correlated random variables.

In Appendix 1, the mean and variance for the binomial distribution are derived. And in Appendix 2, we explain how the binomial distribution can be used in developing the binomial option pricing model.

6.2 Discrete and Continuous Random Variables

A random experiment generally results in numerical values that can be attached to the possible outcomes. In experiments such as throwing a die or measuring a firm's net earnings, the outcomes are naturally in numerical form. The possible outcomes of tossing a fair die are 1, 2, 3, 4, 5, and 6, and the corresponding probabilities are $\frac{1}{6}$ for each outcome, as we saw in Chap. 5. The result of a random experiment can be conveniently described by a random variable. A *random variable* is a variable that assigns a numerical value to each possible outcome of a random experiment. We can think of a random variable as a value or magnitude that changes from occurrence to occurrence in no predictable sequence. A breast cancer screening clinic, for example, has no way of knowing exactly how many women will be screened on any one day. So tomorrow's number of patients is a random variable. For another example, say a company manufactures TV sets that are sometimes defective. Buyers return the defective sets for repair. A variable used to describe the number of TV sets that will be returned before the warranty runs out is a random variable.

Random variables are either *discrete* or *continuous*. *A discrete random variable* is one that can take on a countable number of values; usually it is an integer. The number of claims on an automobile policy in a particular year is a discrete random variable. Another discrete random variable is the number of defective parts produced in a particular run. Here the discrete random variable can take on the values 0, 1, 2, . . ., n. If we let X stand for a discrete random variable, then we can use x to represent one of its possible values. In other words, X is a quantity and x a value. For example, before the results of rolling a fair die are observed, the random variable can be used to denote the outcome. This random variable can assume the

specific values $x = 1$, $x = 2$, …, $x = 6$, and each value has a probability of $\frac{1}{6}$. Other discrete random variables include:

1. The number of bids received in a stock offering: $x = 0, 1, 2, \ldots$
2. The number of customers waiting to be served in a bank at a particular time: $x = 0, 1, 2, \ldots$
3. The number of sales made by a salesperson in a given month: $x = 0, 1, 2, \ldots$
4. The number of people in a sample of 800 who favor a particular presidential candidate: $x = 0, 1, 2, \ldots, 800$

In contrast, a *continuous random variable* can take on an uncountable number of values within an interval. The amount of rainfall in a given area is a continuous random variable. This number can take on an infinite number of values – 8.01 in. of rain is different from 8.012 in.. Measurement may stop at some number of decimal points, but the variable is theoretically continuous. Although it is impossible to attach a probability to the amount of rain equaling exactly 8.012000 … inches, it is possible to give the probability that the amount of rain will be within an interval. Continuous random variables can also represent the amount of time it takes to fill a food order at a restaurant or the length of a bolt used in the production of an automobile. Other continuous random variables appear in the following examples:

1. Let X be the arrival time at an airport between 8:00 and 9:00 a.m.: $8{:}00 \leq x \leq 9{:}00$.
2. For a new residential division, the length of time X from completion until a specified number of houses are sold: $a \leq x \leq b$, for $b > a$.
3. Let Y be the amount of orange juice loaded into a 24-oz bottle in a bottling operation: $0 \leq y \leq 24$.
4. The depth at which a successful natural gas drilling venture first strikes natural gas.
5. The weight of a bag of rice bought in a supermarket.

6.3 Probability Distributions for Discrete Random Variables

6.3.1 Probability Distribution

To analyze a random variable, we must generally know the probability that the variable will take on certain values. The *probability function*, or the *probability distribution*, of a discrete random variable is a systematic listing of all possible values a discrete random variable can take on, along with their respective probabilities. The probability that the random variable X will assume the value x is symbolized by $P(X = x)$ or simply $P(x)$. Note that X is a quantity and x a value. Because a discrete probability function takes nonzero values only at discrete points x, it is sometimes called a *probability mass function*.

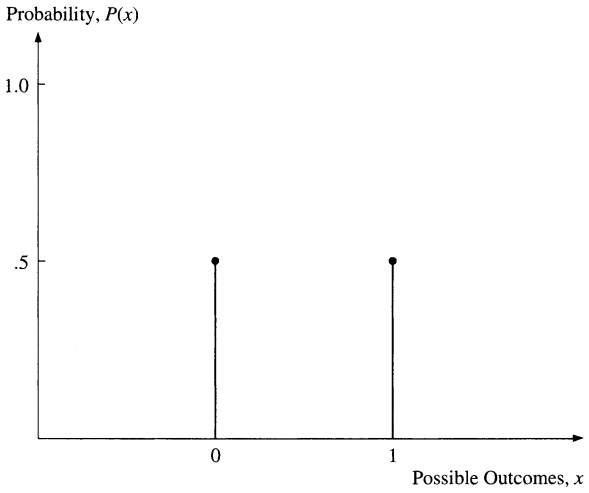


Fig. 6.1 Probability distribution for Example 6.1

Example 6.1 Probability Distribution for the Outcome of Tossing a Fair Coin. Suppose a fair coin is tossed. Let the random variable X represent the outcome, where 1 denotes heads and 0 denotes tails. The probability that heads appears is $P(X = 1) = .5$, and the probability that tails appears is $P(X = 0) = .5$. Figure 6.1 shows a probability distribution where the possible outcomes are charted on the horizontal axis and probabilities on the vertical axis. The spikes in the figure place the probability of heads and that of tails at .5. Note that the probabilities for both outcomes (heads and tails) are between 0 and 1 inclusive and that the sum of both probabilities is 1.

Example 6.2 Probability Distribution for Section Assignment in a Marketing Course. Suppose that five sections of a marketing course are offered and each section has a different number of openings (see Table 6.1). If students are assigned randomly to the sections, then a probability distribution can be drawn for section assignment. The probability that a student is assigned to section 1, $P(X = 1)$, is equal to $\frac{23}{179} = .128$; the probability that a student is assigned to section 2, $P(X = 2)$, is equal to $\frac{45}{179} = .251$. The rest of the probabilities are calculated in the same manner, as indicated in the third column of Table 6.1. Figure 6.2 shows this probability distribution.

Example 6.3 Probability Distribution for the Outcome of Rolling a Fair Die. The probability distribution for the roll of a fair die is shown in Fig. 6.3. Here all the spikes are equal to $\frac{1}{6}$ because $P(X = 1) = P(X = 2) = \cdots = P(X = 6) = \frac{1}{6}$.

Examples 6.1, 6.2, and 6.3 show that the probability of a random variable X taking on the specific value x can be denoted as $P(X = x)$. The probability distribution of a random variable is a representation of the probabilities for *all* possible outcomes. $P(X = x)$ is the probability function of random variable X denoting the

Table 6.1 Probability distribution of marketing course openings

Section, x	Openings	Probability, $P(x)$
1	23	23/179 = .128
2	45	45/179 = .251
3	21	21/179 = .117
4	56	56/179 = .313
5	*34*	*34/179 = .190*
Total	179	1.00

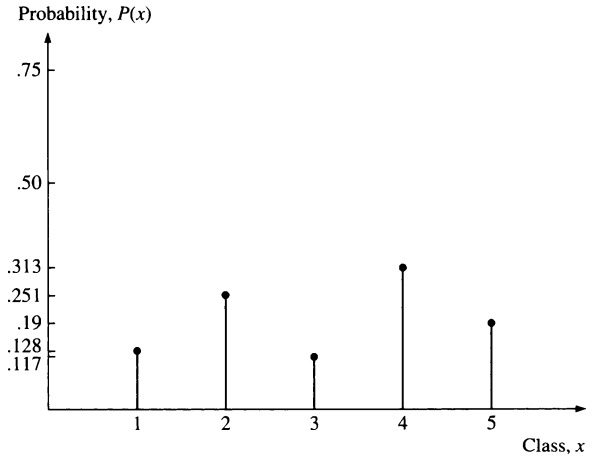


Fig. 6.2 Probability distribution for marketing course openings

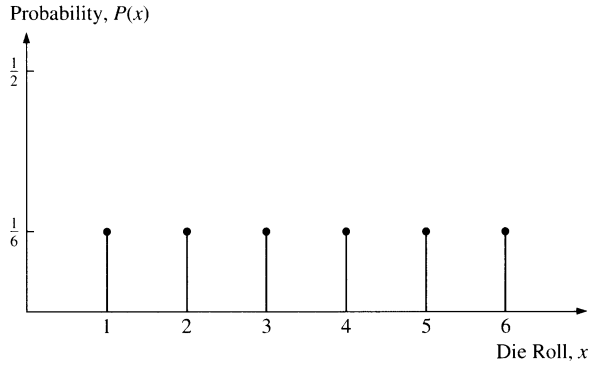


Fig. 6.3 Probability function for Example 6.3

probability that X takes on the value of x. This expression can be rewritten as $P(X = x) = P(x)$ where the function is evaluated at all possible values of X.

For all probability functions of discrete random variables,

1. $P(x_i) \geq 0$ for all i
2. $\sum_{i=1}^{n} P(x_i) = 1$

where x_i is the ith observation of random variable X. Property 1 states that the probabilities cannot be negative. Property 2 implies that the individual probabilities add up to 1.

6.3.2 Probability Function and Cumulative Distribution Function

For some problems, we need to find the probability that X will assume a value less than or equal to a given number. A function representing such probabilities is called a *cumulative distribution function* (cdf) and is usually denoted by $F(x)$. If $x_1, x_2, \ldots, x_m$ are the m values of X given in increasing order (i.e., if $x_1 < x_2 < \ldots, < x_m$), then the cumulative distribution function of x_k, $1 \leq k \leq m$, is given by

$$F(x_k) = P(X \leq x_k) \tag{6.1}$$

In Eq. 6.1, $P(X \leq x_k)$ gives us the probability that X will be less than or equal to x_k. The relationship between the probability function $P(x)$ and the cumulative distribution function $F(x_k)$ can be expressed as follows:

$$F(X_k) = P(x_1) + P(x_2) + \cdots + P(x_k) = \sum_{i=1}^{k} P(x_i) \tag{6.2}$$

Because the values outside the range of X (values smaller than x_1 or larger than x_m) occur only with probability equal to zero, we may equally well write

$$F(x_k) = \sum_{i=-\infty}^{k} P(x_i) \quad \text{for } k \leq m \tag{6.2a}$$

The following examples show how to calculate the cumulative distribution function.

Example 6.4 Cumulative Distribution Function for Rolling a Fair Die. Reviewing Example 6.3, we find that the value of a random variable X and its probability of occurring upon the rolling of a fair die are listed in the first and second columns of Table 6.2, respectively. Because $x_1 < x_2 < \ldots < x_6$, the cumulative distribution function can be calculated in accordance with Eq. 6.1 as follows:

$$F(1) = P(X = 1) = \tfrac{1}{6}$$

$$F(2) = P(X = 1) + P(X = 2) = \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{3}$$

$$F(3) = P(X = 1) + P(X = 2) + P(X = 3) = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2}$$

Table 6.2 Cumulative distribution function for the outcome of tossing a fair die

x	Probability function, $P(x)$	Cumulative distribution function, $F(x)$
1	1/6	1/6
2	1/6	1/3
3	1/6	1/2
4	1/6	2/3
5	1/6	5/6
6	1/6	1
Total	1	

and so on, as listed in the last column of Table 6.2. The cumulative distribution function is shown in Fig. 6.4.

This graph is a *step function*: the values change in discrete "steps" at the indicated integral values of the random variable X. Thus, $F(x)$ takes the value 0 to the left of the point $x = 1$, steps up to $F(x) = \frac{1}{6}$ at $x = 1$, and so on. The dot at the left of each horizontal line segment indicates the probability for that integral value of x. At these points, the values of the cumulative distribution function are read from the upper line segments.

6.4 Expected Value and Variance for Discrete Random Variables

Probability distributions tell us a great deal about the probability characteristics of a random variable. Graphical depictions reveal at a glance the central tendency and dispersion of a discrete distribution, but numerical measures of central tendency and dispersion for a probability distribution are also useful. The mean (central location) of a random variable is called the *expected value* and is denoted by $E(X)$. The expected value of a random variable, which we also denote as μ, is calculated by summing the products of the values of the random variable and their corresponding probabilities:

$$\mu = E(X) = \sum_{i=1}^{N} x_i P(x_i) \tag{6.3}$$

John Kraft, a marketing executive for Computerland, Inc., must decide whether to use a new label on one of the company's personal computer products. The firm will gain $900,000 if Mr. Kraft adopts the new label and it turns out to be superior to the old label. The firm will lose $600,000 if Mr. Kraft adopts the new label and it proves to be inferior to the old one. In addition, Mr. Kraft feels that there is

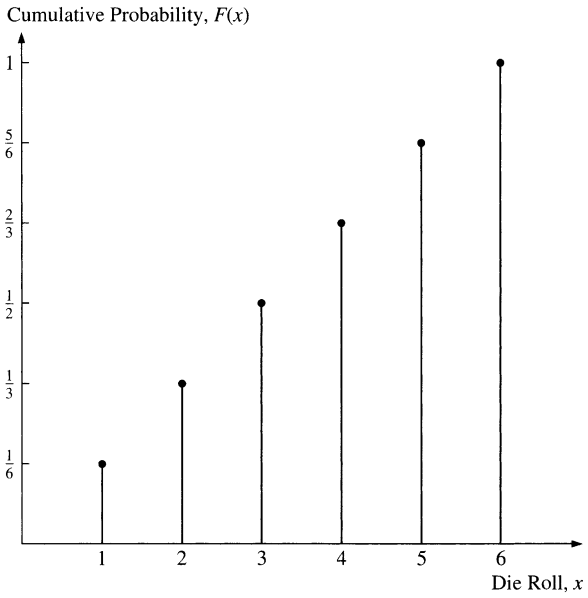


Fig. 6.4 Cumulative probability distribution for Example 6.4

.60 probability that the new label is superior to the old one and .40 probability that it is not. The expected value of the firm's gain for adopting the new label is

$$\begin{aligned} E(X) &= (\$900,000)(.6) + (-\$600,000)(.4) \\ &= \$300,000 \end{aligned}$$

Therefore, Mr. Kraft should consider adopting the new label.

Example 6.5 Expected Value for Earnings per Share. Suppose a stock analyst derives the following probability distribution for the earnings per share (EPS) of a firm.

EPS ($)	$P(x)$	EPS ($)	$P(x)$
1.50	.05	2.25	.15
1.75	.30	2.50	.10
2.00	.35	2.75	.05

To calculate the expected value (the mean of the random variable), we multiply each EPS by its probability and then add the products:

$$\begin{aligned} E(X) &= 1.50(.05) + 1.75(.30) + 2.00(.35) + 2.25(.15) + 2.50(.10) + 2.75(.05) \\ &= 2.025 \end{aligned}$$

The expected value for the earnings per share is 2.025.

Example 6.6 Expected Value of Ages of Students. Suppose the distribution of the ages of students in a class is

Age	$P(x)$	Age	$P(x)$
20	.06	24	.10
21	.10	25	.03
22	.28	26	.04
23	.39		

The expected age is

$$\begin{aligned} E(X) &= 20(.06) + 21(.10) + 22(.28) + 23(.39) + 24(.10) + 25(.03) + 26(.04) \\ &= 22.62 \end{aligned}$$

In addition to calculating expected value for a probability distribution, we can compute the variance and standard deviation as measures of variability. The variance of a distribution is computed similarly to the variance for raw data, which we discussed in Chap. 4. The variance is the summation of the square of the deviations from the mean, multiplied by the corresponding probability:

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 P(x_i) \tag{6.4}$$

where σ^2 is the variance of X, μ is the mean of X, and $P(x_i)$ is the probability function of x_i. If $P(x_1) = P(x_2) = \cdots = P(x_N) = 1/N$, then Eq. 6.4 reduces to

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \tag{6.4a}$$

[1] From Eq. 6.4, the variance of X is

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{N} (x_i^2 - 2\mu x_i + \mu^2) P(x_i) \\ &= \sum_{i=1}^{N} x_i^2 P(x_i) - 2\mu \sum_{i=1}^{N} x_i P(x_i) + \mu^2 \sum_{i=1}^{N} P(x_i) \end{aligned}$$

Because $\sum_{i=1}^{N} x_i P(x_i) = \mu$ and $\sum_{i=1}^{N} P(x_i) = 1$,

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{N} x_i^2 P(x_i) - 2\mu^2 + \mu^2 = \sum_{i=1}^{N} x_i^2 P(x_i) - \mu^2 \\ &= \sum_{i=1}^{N} x_i^2 P(x_i) - \mu^2. \end{aligned}$$

The standard deviation is the square root of the variance. An alternative – and possibly easier – way to calculate the variance is to sum the product of the square of values of the random variables multiplied by the corresponding probabilities and then subtract the expected value squared[1]:

$$\sigma^2 = \sum_{i=1}^{N} x_i^2 P(x_i) - \mu^2 \tag{6.5}$$

Example 6.7 Expected Value and Variance: Defective Tires. Suppose the following table gives the number of defective tires that roll off a production line in a day. Calculate the mean and variance.

Defects	Probability
0	.05
1	.15
2	.20
3	.25
4	.25
5	.10

The expected value is equal to (0)(.05) + (1)(.15) + (2)(.20) + (3)(.25) + (4)(.25) + (5)(.10) = 2.8. Thus, the mean number of defective tires in a production run is 2.8 tires in a day.

The variance is

$$\begin{aligned}&(0-2.8)^2(.05)+(1-2.8)^2(.15)+(2-2.8)^2(.20)\\&+(3-2.8)^2(.25)+(4-2.8)^2(.25)+(5-2.8)^2(.10)=1.86\end{aligned}$$

The alternative formula yields the same answer for the variance:

$$\begin{aligned}&\left[\left(0^2\right)(.05)+\left(1^2\right)(.15)+\left(2^2\right)(.20)+\left(3^2\right)(.25)+\left(4^2\right)(.25)+\left(5^2\right)(.10)\right]\\&-(2.8)^2\\&=1.86\end{aligned}$$

Example 6.8 Expected Value and Variance: Commercial Lending Rate. Returning to the example of commercial lending interest rates in Sect. 5.8, we can tabulate the possible lending rates x and the corresponding probabilities, $P(x)$, as follows:

x	$P(x)$	x	$P(x)$
15 %	.100	18 %	.150
17	.075	11	.100
20	.075	13	.075
13	.200	16	.075
15	.150		

From formulas for the expected value and the variance for discrete random variables, the mean of X is

$$\begin{aligned} E(X) &= \sum_{i=1}^{N} x_i P(x_i) = \mu \\ &= (.100)(.15) + (.075)(.17) + (.075)(.20) + (.200)(.13) \\ &\quad + (.150)(.15) + (.150)(.18) + (.100)(.11) + (.075)(.13) \\ &\quad + (.075)(.16) \\ &= 15.1\% \end{aligned} \tag{6.6}$$

The standard deviation of X can be calculated from

$$\begin{aligned} \sigma &= \left[\sum_{i=1}^{N} (x_i - \mu)^2 P(x_i)\right]^{1/2} \\ &= \Big[(.100)(15 - 15.1)^2 + (.075)(17 - 15.1)^2 + (.075)(20 - 15.1)^2 \\ &\quad + (.200)(13 - 15.1)^2 + (.150)(15 - 15.1)^2 + (.150)(18 - 15.1)^2 \\ &\quad + (.100)(11 - 15.1)^2 (.075)(13 - 15.1)^2 + (.075)(16 - 15.1)^2\Big]^{1/2} \\ &= 2.51\% \end{aligned} \tag{6.7}$$

A bank manager may use this information to make lending decisions, which is discussed in Application 7.4 in Chap. 7.

6.5 The Bernoulli Process and the Binomial Probability Distribution

In this section, we examine first the Bernoulli process and then the binomial probability distribution and its applications.

6.5.1 The Bernoulli Process

The binomial distribution is based on the concept of a *Bernoulli process*, which has three important characteristics. First, a Bernoulli process is a repetitive random process consisting of a series of independent trials. This means that the outcome of one trial does not affect the probability of the outcome of another. Second, only two outcomes are possible in each trial: success or failure. The probability of success is equal to p, and the probability of failure is $(1 - p)$. Third, the probabilities of

success and failure are the same in each trial. For example, suppose the owner of an oil firm believes that the probability of striking oil is .10. Success is defined as striking oil and failure as not striking oil. If the probability of striking oil is .10 on every trial and all the trials are independent of each other, then this is a Bernoulli process. Note that the events of striking oil and not striking oil are mutually exclusive.

A simple example of a Bernoulli process is the tossing of a fair coin. The outcomes can be classified into the events' success (e.g., heads) and failure (tails). The outcomes are mutually exclusive, and the probability of success is constant at .5. The MINITAB output of the Bernoulli process for the first four experiments of Table 5.1 is presented in Fig. 6.5. Columns C1, C2, C3, and C4 present the number and sequence of heads and tails occurring for random experiments with $N = 10$, 20, 30, and 40.

6.5.2 Binomial Distribution

If n trials of a Bernoulli process are observed, then the total number of successes in the n trials is a random variable, and the associated probability distribution is known as a *binomial distribution*. The number of successes, the number of trials, and the probability of success on a trial are the three pieces of information we need to generate a binomial distribution.

To develop the binomial distribution, assume that each of the n trials of an experiment will generate one of two outcomes, a success, S, or a failure, F. Suppose the trials generate x successes and $(n - x)$ failures. The probability of success on a particular trial is p, and the probability of failure is $(1 - p)$. Thus, the probability of obtaining a specific sequence of outcomes is

$$p^x(1-p)^{n-x} \tag{6.8}$$

Equation 6.8 presents the joint probability of x successes and $(n - x)$ failures occurring simultaneously. Because the n trials are independent of each other, the probability of any particular sequency of outcomes is, by the multiplication rule of probabilities (Sect. 5.6), equal to the product of the probabilities for the individual outcomes.

6.5.3 Probability Function

There are several ways in which x successes can be arranged among $(n - x)$ failures. Therefore, the probability of x successes in n trials for a binomial random variable X is

Fig. 6.5 MINITAB output of Bernoulli process for four experiments ($N = 10$, $N = 20$, $N = 30$, and $N = 40$)

```
MTB >  RANDOM 10 C1;
SUBC>  BERNOULI 0.5.
MTB >  RANDOM 20 C2;
SUBC>  BERNOULI 0.5.
MTB >  RANDOM 30 C3;
SUBC>  BERNOUL1 0.5.
MTB >  RANDOM 40 C4;
SUBC>  BERNOULI 0.5.
MTB >  PRINT C1-C4
```

Data Display

Row	C1	C2	C3	C4
1	1	1	1	0
2	0	1	1	0
3	0	1	1	0
4	1	0	1	1
5	0	1	1	1
6	1	1	1	0
7	0	0	0	0
8	0	1	1	0
9	0	0	0	0
10	0	1	1	1
11		0	0	1
12		1	0	1
13		1	1	0
14		0	0	1
15		1	0	1
16		0	1	0
17		0	1	1
18		0	1	1
19		0	0	0
20		1	0	1
21			0	0
22			0	0
23			0	0
24			0	1
25			0	0
26			0	1
27			0	0
28			0	1
29			1	1
30			0	0
31				1
32				0
33				0
34				1
35				1
36				0
37				0
38				0
39				0
40				0

Table 6.3 Probability distribution of JNJ stock 4 days later

Outcome, e		Probability, $p(e)$
e_1	(UUUU)	(.4)(.4)(.4)(.4) = .0256
e_2	(UUUD)	(.4)(.4)(.4)(.6) = .0384
e_3	(UUDU)	(.4)(.4)(.6)(.4) = .0384
e_4	(UUDD)	(.4)(.4)(.6)(.6) = .0576
e_5	(UDUU)	(.4)(.6)(.4)(.4) = .0384
e_6	(UDUD)	(.4)(.6)(.4)(.6) = .0576
e_7	(UDDU)	(.4)(.6)(.6)(.4) = .0576
e_8	(UDDD)	(.4)(.6)(.6)(6) = .0864
e_9	(DUUU)	(.6)(.4)(.4)(.4) = .0384
e_{10}	(DUUD)	(.6)(.4)(.4)(.6) = .0576
e_{11}	(DUDU)	(.6)(.4)(.6)(.4) = .0576
e_{12}	(DUDD)	(.6)(.4)(.6)(.6) = .0864
e_{13}	(DDUU)	(.6)(.6)(.4)(.4) = .0576
e_{14}	(DDUD)	(.6)(.6)(.4)(.6) = .0864
e_{15}	(DDDU)	(.6)(.6)(.6)(.4) = .0864
e_{16}	(DDDD)	(.6)(.6)(.6)(.6) = .1296

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, x = 0, 1, \ldots, n, \tag{6.9}$$

where

$\binom{n}{x}$ = n combinations taken x at a time

$$n! = n(n-1)(n-2)(n-3)\cdots(1)$$

The symbol $n!$ is read "n factorial." When $n = 0$, then $n! = 0! = 1$. Equation 6.9 is the *binomial probability function*, which gives the probability of x successes in n trials: using this formula, we can evaluate a binomial probability.

Example 6.9 Probability Distribution for JNJ Stock. Suppose that the price of a share of stock in Johnson & Johnson company in the future will either go up (U) or come down (D) in 1 day with the probabilities .40 and .60, respectively. Calculate the probability of each possible outcome of the stock price 4 days later.[2]

Using the outcome tree approach discussed in Appendix 1 of Chap. 5, we find the possible outcomes e and probabilities $p(e)$ indicated in Table 6.3.

[2] Assume that the price movement of JNJ stock today is completely independent of its movement in the past. See Example 6.23 in Appendix 2 for further discussion.

The probability of JNJ stock going up three times and coming down once is the sum of the probabilities associated with e_2, e_3, e_5, and e_9: .0384 + .0384 + .0384 + .0384 = .1536.

Alternatively, this probability can be calculated in terms of the binomial combination formula (Eq. 6.9):

$$\binom{4}{3}(.4)^3(.6) = \frac{4!}{(4-3)!3!}(.0384) = .1536$$

Hence, the binomial combination formula can be used to replace the diagram for calculating such a probability.

Example 6.10 Probability Function of Insurance Sales. Assume that an insurance sales agent believes that the probability of she making a sale is .20. She makes five contacts and, eager to leave nothing to chance, calculates a binomial distribution:

$$P(0 \text{ success}) = \frac{5!}{0!5!}.2^0.8^5 = .3277$$

$$P(1 \text{ success}) = \frac{5!}{1!4!}.2^1.8^4 = .4096$$

$$P(2 \text{ successes}) = \frac{5!}{2!3!}.2^2.8^3 = .2048$$

$$P(3 \text{ successes}) = \frac{5!}{3!2!}.2^3.8^2 = .0512$$

$$P(4 \text{ successes}) = \frac{5!}{4!1!}.2^4.8^1 = .0064$$

$$P(5 \text{ successes}) = \frac{5!}{5!0!}.2^5.8^0 = .0003$$

Alternatively these numbers can be calculated by the MINITAB program as shown here:

```
MTB > SET INTO C1
DATA> 0 1 2 3 4 5
DATA> END
MTB > PDF C1;
SUBC> BINOMIAL 5 0.2.
```

Probability Density Function

```
Binomial with n = 5 and p = 0.200000
```

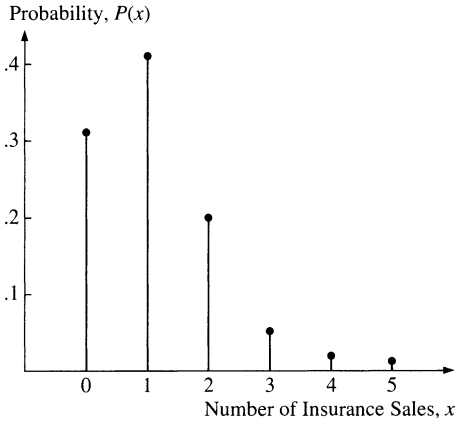


Fig. 6.6 Binomial probability distribution for Example 6.10 ($n = 5, p = .2$)

x	P(X = x)
0.00	0.3277
1.00	0.4096
2.00	0.2048
3.00	0.0512
4.00	0.0064
5.00	0.0003

Figure 6.6 gives the probability distribution for this sales agent's successes. Because the events of the sales agent's number of successes are mutually exclusive, the probability that she has three or more successes is equal to $P(3 \text{ successes}) + P(4 \text{ successes}) + P(5 \text{ successes}) = .0512 + .0064 + .0003 = .0579$.

Example 6.11 Cumulative Probability Distribution for Insurance Sales. Suppose the sales agent we met in Example 6.10 wants to determine the probability of making between 1 and 4 sales:

$$P(1 \text{ success}) + P(2 \text{ successes}) + P(3 \text{ successes}) + P(4 \text{ successes}) = .672$$

Unless the number of trials n is very small, it is easier to determine binomial probabilities by using Table Al in Appendix A of this book. All three variables listed in Eq. 6.9 ($n, p,$ and x) appear in the binomial distribution table extracted from the National Bureau of Standards tables. Using probabilities from this table, we can calculate both individual probabilities and cumulative probabilities.

The individual probabilities drawn for Example 6.11 from the binomial table are listed in Table 6.4. These probabilities are identical to those we found with Eq. 6.9.

The cumulative binomial function can be denned as

$$B(n,p) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \tag{6.10}$$

Table 6.4 Part of binomial table ($n = 5, p = .2$)

x	$P(x)$
5	.0003
4	.0064
3	.0512
2	.2048
1	.4096
0	.3277

Using Table 6.4, we can calculate the cumulative probabilities for the sales agent having two or more successes:

$$\begin{aligned} P(X \geq 2|n = 5, p = .2) &= P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) \\ &= \sum_{x=2}^{5} \binom{n}{x} (.2)^x (.8)^{5-x} \\ &= .2048 + .0512 + .0064 + .0003 = .2627 \end{aligned}$$

In a nationwide poll of 2,052 adults by the American Association of Retired Persons (*USA Today,* August 8, 1985), approximately 40 % of those surveyed described the current version of the federal income tax system as fair. Suppose we randomly sample 20 of the 2,052 adults surveyed and record x as the number who think the federal income tax system is fair. To a reasonable degree of approximation, x is a binomial random variable. The probability that x is less than or equal to 10 can be defined as[3]

$$\begin{aligned} &P(X \leq 10|n = 20, p = 0.4) \\ &= \sum_{x=1}^{10} \binom{n}{x} (0.4)^x (0.6)^{20-x} \\ &= 0 + .005 + .0031 + .0123 + .0350 + .0746 + .1244 \\ &\quad + .1659 + .1797 + .1597 + .1171 = .8725 \end{aligned}$$

Another situation that requires the use of a binomial random variable is *lot acceptance sampling*, where we must decide, on the basis of *sample* information about the quality of the lot, whether to accept a lot (batch) of goods delivered from a manufacturer (see Appendix 1 in Chap. 11 for further detail). It is possible to calculate the probability of accepting a shipment with any given proportion of defectives in accordance with Eq. 6.9.

Example 6.12 Cumulative Probability Distribution: A Shipment of Calculator Chips. A shipment of 800 calculator chips arrives at Century Electronics. The contract specifies that Century will accept this lot if a sample of size 20 drawn from

[3] Refer to Table A1 in Appendix A at the end of the book.

the shipment has no more than one defective chip. What is the probability of accepting the lot by applying this criterion if, in fact, 5 % of the whole lot (40 chips) turns out to be defective? What if 10 % of the lot is defective?

This is a binomial situation where there are $n = 20$ trials and $p =$ the probability of success (chip is defective) $= .05$. The shipment is accepted if the number of defectives is either 0 or 1, so the probability of the shipment being accepted is

$$\begin{aligned} P(\text{shipment accepted}) &= P(X \leq 1) \\ &= P(0) + P(1) \end{aligned}$$

Using Table A1 in Appendix A ($n = 20$, $p = .05$), we obtain $P(0) = .3585$ and $P(1) = .3774$. Hence, the probability that Century Electronics accepts delivery is

$$P(\text{shipment accepted}) = .3585 + .3774 = .7359$$

Similarly, if 10 % of the items in the shipment are defective (i.e., if $p = .10$), then

$$P(\text{shipment accepted}) = .1216 + .2702 = .3918$$

This implies that the higher the proportion of defectives in the shipment, the less likely is acceptance of the delivery. And that's as it should be.

6.5.4 *Mean and Variance*

The *expected value* (mean) of the binomial distribution is simply the number of trials times the probability of a success:

$$\mu = np \tag{6.11}$$

The variance of the binomial distribution is equal to

$$\sigma^2 = np(1 - p) \tag{6.12}$$

Thus the standard deviation of the binomial distribution is $\sqrt{np(1-p)}$. The derivation of Eqs. 6.11 and 6.12 can be found in Appendix 1.

Example 6.13 Probability Distribution of Insurance Sales. In the insurance sales case we discussed in Examples 6.10 and 6.11, the expected number of sales can be calculated in terms of Eq. 6.11 as $np = 5(.20) = 1$. The variance of the distribution can be calculated in terms of Eq. 6.12 as $np(1 - p) = 5(.2)(.8) = .8$. Thus, the expected number of sales by the sales agent is equal to 1, and the standard deviation is $\sqrt{.8} = .894$.

6.6 The Hypergeometric Distribution (Optional)

In the last section, we described the binomial distribution as the appropriate probability distribution for a situation in which the assumptions of a Bernoulli process are met. A major application of the binomial distribution is in the computation of probability for cases where the trials are independent. If the experiment consists of randomly drawing n elements (samples), with replacement, from a set of N elements, then the trials are independent.

In most practical situations, however, sampling is carried out *without replacement,* and the number sampled is not extremely small relative to the total number in the population. For example, when a researcher selects a sample of families in a city to estimate the average income of all families in the city, the sampling units are ordinarily not replaced prior to the selection of subsequent ones. That is, the families are not replaced in the original population and thus are not given an opportunity to appear more than once in the sample. Similarly, when a sample of accounts receivable is drawn from a firm's accounting records for a sample audit, sampling units are ordinarily not replaced before the selection of subsequent units. Sampling without replacement also takes place in quality control sampling and other sampling. Furthermore, if the number sampled is extremely small relative to the total number of items, then the trial is almost independent even if the sampling is without replacement (as in Example 6.12).[4] Under such circumstances and in sampling with replacement, the binomial distribution can be used in the analysis.

The *hypergeometric distribution* is the appropriate model for sampling without replacement. To solve the following hypergeometric problems, let's divide our population (such as a group of people) into two categories: adults and children. For a population of size N, h members are S (successes) and $(N - h)$ members are F (failures). Let sample size $= n$ trials, obtained without replacement. Let $x =$ number of successes out of n trials (a hypergeometric random variable).

Suppose there are $h = 60$ adults and $N - h = 40$ children. Thus, there are $N = 100$ persons. Numbers from 1 to 100 are assigned to these individuals and printed on identical disks, which are placed in a box. If 10 chips are randomly drawn from the box, then the hypergeometric problem involves calculating the probability of there being x adults and $(n - x)$ children in a sample of size 10. If $n = 10$ people are selected, what is the probability that exactly four adults will be included in the sample?

Because there are $h = 60$ adults, there are $\binom{h}{x}$ possible ways of selecting $x = 4$ adults. Of the $n = 10$ people, $n - x = 10 - 4 = 6$ are children. Hence,

[4] In that case, the probability on the first trial, P, is 50/800 = .0625. On the second trial, p is either 50/799 = .06258 (if the first chip was not defective) or 49/799 = .06133 (if the first chip was defective).

there are $\binom{N-h}{n-x}$ possible ways of selecting $n - x = 6$ children. Thus, the total number of ways of selecting a group of 10 persons that includes exactly four adults and six children is $\binom{h}{x}\binom{N-h}{n-x}$. There are $\binom{N}{n} = \binom{100}{10}$ possible ways of selecting 10 persons from 100 persons. Thus, the probability of selecting a group of 10 persons that includes $x = 4$ adults is

$$\frac{\binom{60}{4}\binom{40}{10-4}}{\binom{100}{10}}$$

6.6.1 The Hypergeometric Formula

From the example we just outlined, we can state the general hypergeometric probability function for a hypergeometric variable X as

$$P(X = x) = P[(x \text{ successes and} (n - x) \text{ failures})] = \frac{\binom{h}{x}\binom{N-h}{n-x}}{\binom{N}{n}} \tag{6.13}$$

The *hypergeometric formula* gives the probability of x successes when a random sample of n is drawn without replacement from a population of N within which h units have the characteristic denoting success. The number of successes achieved under these circumstances is the *hypergeometric random variable*.

Example 6.14 Sampling Probability Function of Party Membership. Consider a group of 10 students in which four are Democrats and six are Republicans. A sample of size six has been selected. What is the probability that there will be only 1 Democrat in this sample?

Using the hypergeometric probability function shown in Eq. 6.13, we have

$$P(x = 1 \text{ and } n - x = 5) = \frac{\binom{4}{1}\binom{10-4}{6-1}}{\binom{10}{6}} = \frac{\frac{4!}{1!(4-1)!}\frac{6!}{5!(6-5)!}}{\frac{10!}{6!4!}}$$

$$= \frac{\frac{24}{(1)(6)}\frac{720}{(120)(1)}}{\frac{3,628,800}{(720)(24)}} = \frac{(4)(6)}{210} = .1143$$

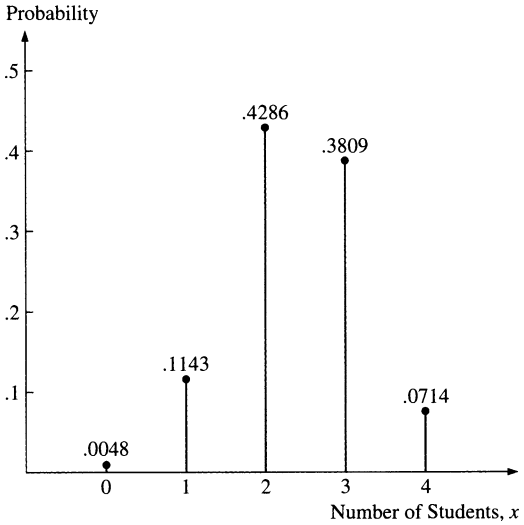


Fig. 6.7 Probability distribution for Example 6.14 (hypergeometric distribution for $N = 10$, $x = 4$, $n = 6$)

Similarly, we can calculate other probabilities. All possible probabilities are as follows:

$P(x = 0) = .0048$
$P(x = 1) = .1143$
$P(x = 2) = .4286$
$P(x = 3) = .3809$
$P(x = 4) = .0714$

The hypergeometric probability distribution is shown in Fig. 6.7.

6.6.2 Mean and Variance

The mean of the hypergeometric probability distribution for Example 6.14 can be calculated by using Eq. 6.3:

$$\mu = \sum_{i=1}^{n} x_i P(x_i) = 0(.0048) + 1(.1143) + 2(.4286) + 3(.3809) + 4(.0714)$$
$$= 2.40$$

On average, we expect 2.40 students to be Democrats. Alternatively, it can be shown that the mean of this distribution is

$$\mu = n(h/N) \tag{6.14}$$

The ratio h/N *is* the proportion of successes on the first trial. The product $n(h/N)$ is similar to the mean of the binomial distribution, np. It can be shown that the *variance of the hypergeometric distribution* is equal to

$$\sigma^2 = \left(\frac{N-n}{N-1}\right)\left[n\left(\frac{h}{N}\right)\left(1-\frac{h}{N}\right)\right] \tag{6.15}$$

In other words, the variance of the hypergeometric distribution is the variance of the binomial distribution with an adjustment factor, $\left(\frac{N-n}{N-1}\right)$. If the sample size is small relative to the total number of objects N, then $\left(\frac{N-n}{N-1}\right)$ is very close to 1. Consequently, the binomial distribution can be used to replace the hypergeometric distribution.[5]

Example 6.15 Mean and Variance of a Hypergeometric Probability Function. Using the data of Example 6.14, we can calculate the mean and variance of a hypergeometric function as follows:

$$\mu = 6\left(\frac{4}{10}\right) = 2.4$$

$$\sigma^2 = \left(\frac{10-6}{10-1}\right)\left[(6)\left(\frac{4}{10}\right)\left(1-\frac{4}{10}\right)\right]$$
$$= .64$$

6.7 The Poisson Distribution and Its Approximation to the Binomial Distribution

In the previous two sections, we have discussed two major types of discrete probability distributions, one for binomial random variables and the other for hypergeometric random variables. Both of these random variables were defined in terms of the number of success, and these successes were *obtained within a fixed number of trials* of some random experiment. In this section, we will discuss a distribution called the *Poisson distribution*. This distribution can be used to deal with a single type of outcome or "event," such as number of telephone calls that come through a switchboard and number of accidents. It is also possible to use a Poisson distribution to investigate the probability of, say, a certain number of defective parts in a plant in a 1-year period, a certain number of sales in a given week, and a certain number of customers entering a bank in a day.

[5] The approximation is valid only when N is large. Usually we require $N/n \geq 20$.

6.7.1 The Poisson Distribution

The *Poisson distribution*, which is named after the French mathematician Simeon Poisson, is useful for determining the probability that a particular event will occur a certain number of times over a specified period of time or within the space of a particular interval. For example, the number of customer arrivals per hour at a bank or other servicing facility is a random variable with Poisson distribution. Here are some other random variables that may exhibit a Poisson distribution:

1. The number of days in a given year in which a 50-point change occurs in the Dow Jones Industrial Average
2. The number of defects detected each day by a quality control inspector in a light bulb plant
3. The number of breakdowns per month that a supercomputer experiences
4. The number of car accidents that occur per month (or week or day) in the city of Princeton, New Jersey

The formula for the Poisson probability distribution is

$$P(X = x) = e^{-\lambda}\lambda^{x}/x! \text{ for } x = 0, 1, 2, 3, \ldots \text{ and } \lambda > 0 \tag{6.16}$$

where X represents the discrete Poisson random variable; x represents the number of rare events in a unit of time, space, or volume; λ is the mean value of x; e is the base of natural logarithms and is approximately equal to 2.71828; and ! is the factorial symbol.

It can be shown that the value of both the mean and the variance of a Poisson random variable X *is* λ. That is,

$$E(X) = \lambda \tag{6.17a}$$

$$\text{Var}(X) = \lambda \tag{6.17b}$$

We will explore this distribution further in Chap. 9 when we discuss the exponential distribution.

In studying a retailer's supply account at a large US Air Force base, the Poisson probability distribution was used to describe the number of customers (x) in a 7-day lead time period *(Management Science,* April 1983). Here "lead time" is used to describe the time needed to replenish a stock item.

Items were divided into two categories for individual analysis. The first category was items costing $5 or less and the second category was items costing more than $5. The mean number of customers during lead time for the first category was estimated to be .09. For the second category, the mean was estimated to be .15.

From Eqs. 6.17a and b, the mean and variance for the number x of customers who demand items that cost over $5 during lead time is

Table 6.5 Probability function for Example 6.16 ($\lambda = 5$)

x	$P(x)$
4	.1755
3	.1404
2	.0842
1	.0337
0	.0067

$$E(X) = \text{Var}(X) = \lambda = .15.$$

From Eq. 6.16, the probability that no customers will demand an item that costs $5 or less during the lead time is

$$P(X = 0) = e^{-0.09}(.09)^0/0! = .9139$$

Example 6.16 Customer Arrivals in a Bank. Suppose the average number of customers entering a bank in a 30-min period is five. The bank wants to determine the probability that four customers enter the bank in a 30-min period. Substituting $\lambda = 5$ and $X = 4$ into Eq. 6.16, we obtain

$$P(X = 4) = \left(e^{-5}\right)\left(5^4\right)/4!$$

Table A2 in Appendix A of this book is a Poisson probability table that can be used to calculate probabilities. From this table, we find that $(e^{-5})(5^4)/4! = .1755$. As another example, say we know that the probability that three customers enter the bank is .1404. Using the Poisson probability table, we can calculate the other individual probabilities for $X = 0$, 1, and 2. Table 6.5 gives the probability function for $X = 0, 1, \ldots, 4$.

Our calculations can tell us such things as the probability that 0, 1, 2, 3, or 4 customers arrive within a 20-min period. From Table 6.5, we know that the probability that four or fewer individuals enter the bank is $.1755 + .1404 + .0842 + .0337 + .0067 = .4405$.

We could continue by calculating the probabilities for more than four customers and eventually produce a Poisson probability distribution for this bank. Table 6.6 shows such a distribution. To produce this table, we used Eq. 6.16. The probability of more than four customer arrivals can also be calculated from Table A2.

Alternatively, MINITAB can be used to calculate part of Table 6.6 as follows:

Table 6.6 Poisson probability distribution of customer arrivals per 3-min period

x = number of customer arrivals	$P(x)$ = probability of exactly that number
0	.0067
1	.0337
2	.0842
3	.1404
4	.1755
5	.1755
6	.1462
7	.1044
8	.0653
9	.0363
	.9682
10 or more	.0318 (1 – .9682)
	1.0000

```
MTB > SET INTO C1
DATA> 0 1 2 3 4 5 6 7
DATA> END
MTB > PDF C1;
SUBC> POISSON 5.
         K        P(X = K)
      0.00         0.0067
      1.00         0.0337
      2.00         0.0842
      3.00         0.1404
      4.00         0.1755
      5.00         0.1755
      6.00         0.1462
      7.00         0.1044
MTB > PAPER
```

Figure 6.8 uses MINITAB to illustrate graphically the Poisson probability distribution of the number of customer arrivals.

Example 6.17 Defective Spark Plug. In one day's work on a spark plug assembly line, the average number of defective parts is 2. The manager is concerned that more than four defectives could occur and wants to estimate the probability of that happening. Using the Poisson distribution table (Table A2 in Appendix A), she determines that the probability of 0–4 defective spark plugs is $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = .1353 + .2707 + .2707 + .1804 + .0902 = .9473$. Then the probability of having more than four defective spark plugs is $1 - .9473 = .0527$.

```
MTB > SET INTO C1
DATA> 0123456789
DATA> END
MTB > SET INTO C2
DATA> 0.0067 0.0337 0.0842 0.1404 0.1775 0.1775 0.1462 0.1044 0.0653
0.0363
DATA> END
MTB > GPRO
* NOTE  *  Professional Graphics are enabled.
           Standard Graphics are disabled.
           Use the GSTD command to enable Standard Graphics.
MTB >  Plot C2*C1;
SUBC>   Project;
SUBC>   Axis 1;
SUBC>     Label "X";
SUBC>   AXIS 2;
SUBC>     Label "PROBABILITY".
```

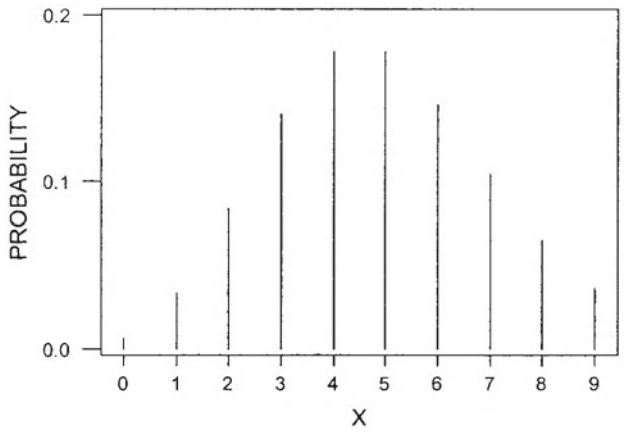


Fig. 6.8 *MINITAB* output of Poisson probability distribution for the number of customer arrivals

6.7.2 The Poisson Approximation to the Binomial Distribution

The Poisson distribution can sometimes be used to *approximate* the binomial distribution and avoid tedious calculations.

If the number of trials in a binomial, n, is large, then a Poisson random variable with $\lambda = np$ will provide a reasonable approximation. This is a good approximation, provided that n is large ($n > 20$) and p is small ($p < .05$):

$$P(X = x) = \frac{e^{-np}(np)^x}{x!} \tag{6.18}$$

Example 6.18 Comparison of the Poisson and Binomial Probability Approaches. Suppose 20 parts are selected from a production process and tested for defects. The manager of the firm wants to determine the probability that three defectives are

encountered. Previous experience indicates that the probability of a part being defective is .05. The mean is $np = (20)(.05) = 1$. Setting $\lambda = 1$, we can use the Poisson distribution formula of Eq. 6.18 to calculate the probability:

$$P(X = 3) = 1^3 e^{-1}/3! = .0613$$

If we use the binomial distribution formula of Eq. 6.9, then the probability is

$$P(x) = \frac{20!}{3!(20-3)!}(.05)^3(.95)^{17} = .0596$$

The difference between .0613 and .05916 is slight (only about .2 %).

6.8 Jointly Distributed Discrete Random Variables (Optional)

In Sects. 5.4 and 5.5, we discussed conditional, joint, and marginal probabilities in terms of events. We now consider these probabilities for two or more related discrete random variables. For a single random variable, the probabilities for all possible outcomes can be summarized by using a probability function; for two or more possible related discrete random variables, the probability function must define the probabilities that the random variables of interest simultaneously take specific values.

6.8.1 Joint Probability Function

Suppose we want to know the probability of a worker being a member of a labor union *and* over age 50. We now concern ourselves with the distribution of random variables, age (X) and membership in a labor union (Y). In notation, the probability that X takes on a value x and that Y takes on a value y is given by

$$P(x, y) = P(X = x, Y = y) \tag{6.19}$$

Equation 6.19 represents the *joint probability function* of X and Y. Joint probabilities are usually presented in tabular form so that the probabilities can be identified easily. *Joint probability distributions* of discrete random variables are probability distributions of two or more discrete random variables. The next example illustrates the use of Eq. 6.19.

Example 6.19 Joint Probability Distribution for 100 Students Classified by Sex and by Number of Accounting Courses Taken. Table 6.7 shows the probability function for two random variables, X (the total number of accounting courses a student takes) and Y (the sex of the student, where 1 denotes a male student and 0 a female). The values in the cells of Table 6.7 are joint probabilities of the outcomes denoted by the column and row headings for X and Y. Also displayed in the margins of the table are separate univariate probability distributions of X and Y.

Table 6.7 Joint probability distribution for 100 students classified by sex and number of accounting courses taken

	X				
Y	*2*	*3*	*4*	*5*	Total
0	.14	.17	.08	.12	.51
1	.16	.20	.12	.01	.49
Total	.30	.37	.20	.13	1

The joint probability that a student is female and takes three courses, $P(3, 0) = P(X = 3, Y = 0)$, is equal to .17. The probability that a student is male and takes four courses, $P(4, 1) = P(X = 4, Y = 1)$, is equal to .12. The probabilities inside the box are all joint probabilities, which are, again, probabilities of the intersection of two events.

The probability distribution of a single discrete random variable is graphed by displaying the value of the random variable along the horizontal axis and the corresponding probability along the vertical axis. In the case of a bivariate distribution, two axes are required for the values of random variables and a third for the probability. A graph of the joint probability of Table 6.7 is shown in Fig. 6.9.

6.8.2 Marginal Probability Function

The *marginal probability* can be obtained by summing all the joint probabilities over all possible values. In other words, the probabilities in the margins of the table are the marginal probabilities. These probabilities form marginal probability functions. For example (see Table 6.7 in Example 6.19), the probability that a randomly selected student is female, $P(Y = 0)$, is found by adding the respective probabilities that a female student takes two courses (.14), three courses (.17), four courses (.08), and five courses (.12), for a total of .51. The probability that a randomly selected student is male, $P(Y = 1)$, is therefore .49 (1 − .51). Similarly, the probability that a randomly selected student takes two courses, $P(X = 2)$, is equal to the probability that a female takes two courses (.14), plus the probability that a male takes two courses (.16), for a total of .30. Note that the sum of the marginal probabilities is 1. From these results, we can define *marginal probability functions* for X and Y as follows:

$$P(x_i) = \sum_{j=1}^{m} P(x_i, y_j), \quad i = 1, \ldots, n \tag{6.20}$$

$$P(y_j) = \sum_{i=1}^{n} P(x_i, y_j), \quad j = 1, \ldots, m \tag{6.21}$$

where

x_i = the ith observation of the X variable
y_j = the jth observation of the Y variable

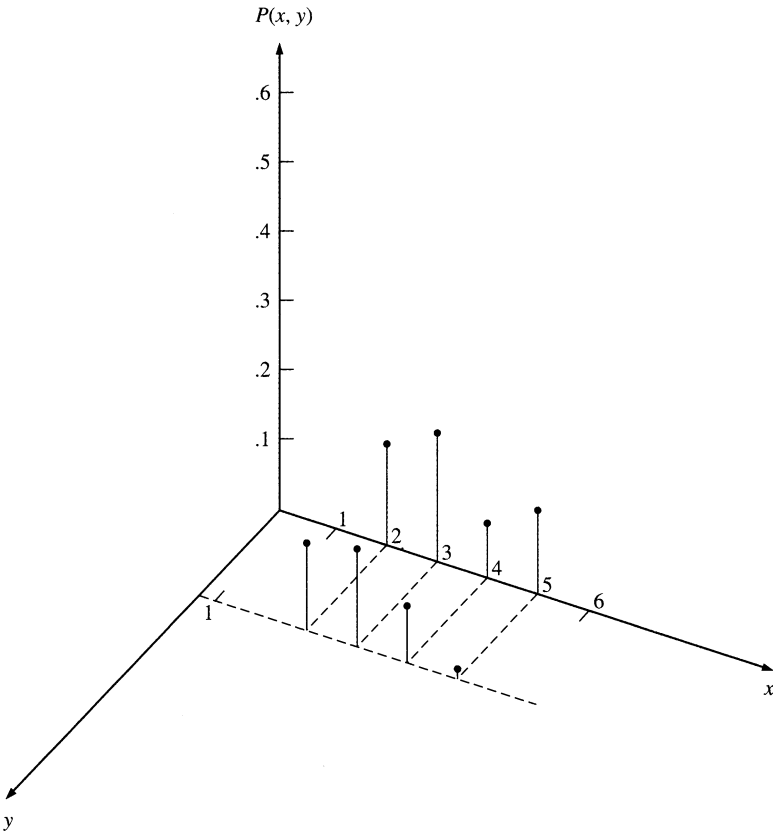


Fig. 6.9 Graph of the bivariate probability distribution shown in Table 6.7

6.8.3 Conditional Probability Function

Conditional probability functions can be calculated from joint probabilities. The conditional probability function of X, given $Y = y$, is

$$P(x|y) = P(x, y)/P(y) \tag{6.22}$$

The conditional probability is found by taking the intersection of the probability of $X = x$ and $Y = y$ and dividing by the probability of Y. For example, the probability from Table 6.7 that a student takes four courses, given that the student is female, is $P(4|0) = P(4,0)/P(0) = .08/.51 = .16$.

Similarly, the probability that a student is male, given that the student takes three courses, is $P(1|3) = P(3,1)/P(3) = .20/.37 = .54$. The *conditional probability distribution* for X given $Y = 1$ is shown in Table 6.8.

Table 6.8 Conditional probability distribution for numbers of accounting courses, given that the student is male (Y = 1)

x	$P(X = x\backslash Y = 1)$
2	.16/.49 = .3265
3	.20/.49 = .4082
4	.12/.49 = .2449
5	.01/.49 = .0204
	1.0000

Table 6.9 Joint probability distribution for consumer satisfaction (x) and number of years of residence in a particular town (Y)

	x				
y	*1*	*2*	*3*	*4*	Total
1	.04	.14	.23	.07	.48
2	.07	.17	.23	.05	.52
Total	.11	.31	.46	.12	1

6.8.4 Independence

Returning to the terminology of events explained in Chap. 5, we saw in Sect. 5.6 that if two events are statistically independent, then $P(B/A) = P(B)$ and $P(B \cap A) = P(B)P(A)$. In random variable notation, the analogous statement is that if X and Y are independent random variables, then

$$\begin{aligned} P(X = x|Y = y) &= P(X = x) \quad \text{for all } X \\ P(Y = y|X = x) &= P(Y = y) \quad \text{and all } Y \end{aligned} \tag{6.23}$$

Equation 6.23 implies that the conditional probability function of X given Y or of Y given X is the same as the marginal probability of X or Y. We will illustrate this definition of independence by returning to Table 6.7.

Suppose we consider the outcome pair (3, 1) – that is, $X = 3$ and $Y = 1$. In this case,

$$P(X = 3|Y = 1) = \frac{.20}{.49} = .4082$$

and

$$P(X = 3) = .37$$

Because $P(X = 3/Y = 1)$ is not equal to $P(X = 3)$, X and Y are not independent.

Example 6.20 Store Satisfaction. Table 6.9 shows the probability function for two random variables: X, which measures a consumer's satisfaction with food stores in a particular town, and Y, the number of years the consumer has resided in that town.[6]

[6] This example is based on the material discussed in J. H. Miller, "Store Satisfaction and Aspiration Theory," *Journal of Retailing*, 52 (Fall 1976), 65–84.

Table 6.10 Conditional probability distribution for satisfaction level for a consumer who has lived in town 6 or more years

x	$P(X = x \backslash Y = 2)$
1	.07/.52 = .1346
2	.17/.52 = .3269
3	.23/.52 = .4423
4	.05/.52 = .0962
	1.0000

Suppose X can take on the value 1, 2, 3, and 4, which reflect a satisfaction level ranging from low to high, and that Y takes on the value 1 if the consumer has lived in the town fewer than 6 years and 2 otherwise. The values in the cells of Table 6.9 are joint probabilities of the respective joint events denoted by the column and row headings for x and y. Also displayed in the margins of the table are separate univariate probability distributions of x and y.

The joint probability that a consumer has satisfaction level 3 and has lived in town fewer than 6 years, $P(3, 1) = P(X = 3, Y = 1)$, is .23. The probability that a consumer has satisfaction level 4 and has lived in the town more than 6 years, $P(4, 2)$, is .05. The probabilities inside the box are all joint probabilities, which are, again, the intersections of two events.

The marginal probability is obtained by summing the joint probabilities over all possible values, as discussed in Example 6.19. For example (see Table 6.9), the probability that a consumer has lived in town fewer than 6 years, $P(Y = 1)$, is found by adding the probabilities that a consumer has satisfaction level 1, 2, 3, and 4, a total of .48. The marginal probability that a consumer has lived in town 6 or more years, $P(Y = 2)$, is therefore .52. Similarly, the probability that a randomly selected consumer has satisfaction level 1, $P(X = 1)$, is equal to the probability that a consumer has lived in town fewer than 6 years, .04, plus the probability that a consumer has lived in the town more than 6 years, .07, for a total of. 11. Note that the sum of the marginal probabilities is equal to 1.

Conditional probability functions can be calculated from joint probabilities as discussed in Example 6.19. The conditional probability is found by taking the intersection of the probability of $X = x$ and $Y = y$ and dividing by the probability of $Y = y$. For example, the probability (from Table 6.9) that a consumer has satisfaction level 4, given that the consumer has lived in town fewer than 6 years, is $P(X = 4/Y = 1) = P(X = 4, Y = 1)/P(Y = 1) = .07/.48 = .1458$.

Similarly, the probability that a consumer has lived in town 6 or more years, given that the consumer has satisfaction level 3, is $P(Y = 2|X = 3) = P(Y = 2, X = 3)/P(X = 3) = .23/.46 = .5$. The conditional probability distribution for X given $Y = 2$ is shown in Table 6.10.

6.9 Expected Value and Variance of the Sum of Random Variables (Optional)

6.9.1 Covariance and Coefficient of Correlation Between Two Random Variables

The concept of expected value and variance of discrete random variables discussed in Sect. 6.4 can be extended to measure the degree of relationship between two discrete random variables X and Y. Here we will discuss two alternative means of determining the possibility of a linear association between two random variables X and Y. These two measures are covariance and coefficient of correlation.

The *covariance* is a statistical measure of the linear association between two random variables X and Y. Its sign reflects the direction of the linear association. The covariance is positive if the variables tend to move in the same direction. If the variables tend to move in opposite directions, the covariance is negative. Specifically, the covariance between X and can be defined as

$$Cov(X,Y) = \sigma_{X,Y} = E[(X-\mu_X)(Y-\mu_Y)] \tag{6.24}$$

where μ_X and μ_y are the means of X and Y, respectively. For discrete variables, Eq. 6.24 can be defined as

$$\sigma_{X,Y} = \sum_{j=1}^{m}\sum_{i=1}^{n}(X_i-\mu_X)\left(Y_j-\mu_Y\right)P\left(X_i,Y_j\right) \tag{6.25}$$

Equation 6.25 can be written as a shortcut formula as follows[7]:

$$\begin{aligned} Cov(X,Y) &= E(XY)-\mu_X\mu_Y \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n}\left(X_iY_j\right)P\left(X_i,Y_j\right) - \left[\sum_{i=1}^{n}X_iP(X_i)\right]\left[\sum_{j=1}^{m}Y_jP\left(Y_j\right)\right] \end{aligned} \tag{6.26}$$

To illustrate, we evaluate the covariance between number of years of residence in the town and satisfaction level, as discussed in Example 6.20. Using the probabilities in Table 6.9, we calculate μ_X, μ_y, and $E(XY)$ as

---

7

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E(XY-Y\mu_X-X\mu_Y+\mu_X\mu_Y) \\ &= E(XY)-\mu_XE(Y)-\mu_YE(X)+\mu_X\mu_Y \\ &= E(XY)-\mu_X\mu_Y-\mu_Y\mu_X+\mu_X\mu_Y = E(XY)-\mu_X\mu_Y \end{aligned}$$

$$\mu_X = \sum_{i=1}^{4} X_i P(X_i) = 1(.11) + 2(.31) + 3(.46) + 4(.12) = 2.59$$

$$\mu_Y = \sum_{j=1}^{2} Y_j P(Y_j) = 1(.48) + 2(.52) = 1.52$$

$$\begin{aligned} E(XY) &= \sum_{j=1}^{m} \sum_{i=1}^{n} (X_i Y_j) P(X_i Y_j) \\ &= (1)(1)(.04) + (1)(2)(.14) + (1)(3)(.23) + (1)(4)(.07) + (2)(1)(.07) \\ &\quad + (2)(2)(.17) + (2)(3)(.23) + (2)(4)(.05) \\ &= 3.89 \end{aligned}$$

Substituting this information into Eq. 6.26, we obtain the covariance:

$$\text{Cov}(X, Y) = 3.89 - (2.59)(1.52) = -0.05$$

The negative value of covariance indicates some tendency toward a negative relationship between number of years of residence in the town and level of satisfaction.

In addition to the direction of the relationship between variables, we may want to measure its strength. We can easily do so by *scaling* the covariance to obtain the coefficient of correlation.

The coefficient of correlation p between X and Y is equal to the covariance divided by the product of the variables' standard deviations. That is,

$$\rho = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{6.27}$$

where ρ = coefficient of correlation, σ_x = standard deviation of X, and σ_Y = standard deviation of Y.

It can be shown that ρ is always less than or equal to 1.0 and greater than or equal to -10:

$$-1 \leq \rho \leq 1$$

Again, let us use data given in Table 6.9 to show how to calculate the correlation coefficient between X and Y. We use Eq. 6.5 to calculate the variances of X and Y:

$$\begin{aligned} \sigma_X^2 &= \sum_{i=1}^{4} X_i^2 P(X_i) - (\mu_X)^2 \\ &= (1)^2(.11) + (2)^2(.31) + (3)^2(.46) + (4)^2(.12) - (2.59)^2 \\ &= .7019 \end{aligned}$$

$$\begin{aligned}\sigma_Y^2 &= \sum_{j=1}^{2} Y_j^2 P(Y_j) - (\mu_Y)^2 \\ &= (1)^2(.48) + (2)^2(.52) - (1.52)^2 \\ &= .2496\end{aligned}$$

Then we substitute $\sigma_{X,Y} = -.15$, $\sigma_X = \sqrt{.7019} = .8378$, and $\sigma_Y = \sqrt{.2496} = .5$ into Eq. 6.27. We obtain

$$\rho = \frac{-.05}{(.8378)(.5)} = -.1194$$

This means the relationship between X and Y is negative, as indicated by the covariance.

As might be expected, the notions of covariance (and coefficient of correlation) and statistical independence are not unrelated. However, the precise relationship between these notions is beyond the scope of this book. Covariance and coefficient of correlation will be discussed in detail in Chaps. 13 and 14.

6.9.2 *Expected Value and Variance of the Summation of Random Variables X and Y*

If X and Y are a pair of random variables with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2 and the covariance between X and Y is $\text{Cov}(X, Y) = \sigma_{x,y}$, then:

1. The expected value of their sum (difference) is the sum (difference) of their expected values:

$$\begin{aligned}E(X + Y) &= \mu_X + \mu_Y \\ E(X - Y) &= \mu_X - \mu_Y\end{aligned}$$

2. The variance of the sum of X and Y, $\text{Var}(X + Y)$, or the difference of X and Y, $\text{Var}(X - Y)$, is the sum of their variances plus (minus) two times the covariance between X and Y:

$$\begin{aligned}\text{Var}(X + Y) &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{X,Y} \\ \text{Var}(X - Y) &= \sigma_X^2 + \sigma_Y^2 - 2\sigma_{X,Y}\end{aligned}$$

Example 6.21 Rates of Return and Variance for a Portfolio. Rates of return for stocks A and B are listed in Table 6.11. Let X = rates of return for stock A, and let Y = rates of return for stock B. The worksheet for calculating μ_X, μ_Y, σ_X, σ_Y, ρ_{XY}, $E(X + Y)$, and $\text{Var}(X + Y)$ is presented in Table 6.12.

Table 6.11 Rates of return for stocks A and B

Time period	Stock A	Stock B
1	.10	−.10
2	−.05	.05
3	.15	.00
4	.05	−.10
5	.00	.10

Table 6.12 Worksheet to calculate summary statistics

Time period	X_i	Y_i	$(X_i - \mu_X)$	$(Y_i - \mu_Y)$	$(X_i - \mu_X)^2$	$(Y_i - \mu_Y)^2$	$(X_i - \mu_X)(Y_i - \mu_Y)$
1	.10	−.10	.05	−.09	.0025	.0081	−.0045
2	−.05	.05	−.10	.06	.010	.0036	−.006
3	.15	.00	.10	.01	.010	.0001	.001
4	.05	−.10	.00	−.09	.00	.0081	.00
5	.00	.10	−.05	.11	.0025	.0121	−.0055
Total	.25	−.05	0	0	.0250	.032	−.015

$$\mu_X = \frac{.25}{5} = .05$$

$$\mu_Y = \frac{-.05}{5} = -.01$$

Substituting information into related formulas for variance, covariance, and correlation coefficient yields

$$\sigma_X^2 = .025/4 = .00625$$

$$\sigma_Y^2 = .032/4 = .008$$

$$\sigma_{X,Y} = -.015/4 = -.00375$$

$$\rho_{X,Y} = \frac{-.00375}{\sqrt{(.00625)(.008)}} = -.5303$$

The MINITAB output of these empirical results is presented in Fig. 6.10. To calculate the expected rate of return and the variance of a portfolio which composes W_1 percent of stock A and W_2 percent of stock B, we need to modify Eqs. 6.28 and 6.29 as[8]

[8] See Appendix 1 in Chap. 13 for further discussion about how to obtain optimal weights for a portfolio.

Fig. 6.10 MINITAB output for Example 6.21

```
MTB   > READ  C1-C2
DATA>    .10  -0.10
DATA>  -0.05   0.05
DATA>   0.15   0.00
DATA>   0.05  -0.10
DATA>   0.00   0.10
DATA>   END
        5   rows   read.
MTB   >   PRINT  C1-C2
```

Data Display

Row	C1	C2
1	0.10	–0.10
2	–0.05	0.05
3	0.15	0.00
4	0.05	–0.10
5	0.00	0.10

```
MTB  >  MEAN  C1
```

Column Mean

```
    Mean of C1 = 0.050000
MTB  >  MEAN  C2
```

Column Mean

```
   Mean of C2 = -0.010000
MTB > STDEV OF C1 PUT INTO K1
```

Column Standard Deviation

```
   Standard deviation of C1 = 0.079057MTB > STDEV
OF C2 PUT INTO K2
```

Column Standard Deviation

```
   Standard deviation of C2 = 0.089443
MTB > LET K3=K1**2
MTB > LET K4=K2**2
MTB > PRINT K3-K4
```

Data Display

```
K3    0.00625000
K4    0.00800000
MTB > COVARIANCE C1 C2
```

Covariances

```
                    C1              C2
C1         0.00625000
C2        -0.00375000      0.00800000

MTB > CORRELATION C1 C2
```

Correlations (Pearson)

Correlation of C1 and C2 = –0.530

$$Var(Rp) = Var(W_1X + W_2Y)$$
$$= W_1^2\ Var(X) + W_2^2\,\mathrm{Var}(Y) + 2W_1W_2Cov(X,Y) \tag{6.28}$$

$$E(Rp) = E(W_1X + W_2Y) = W_1E(X) + W_2E(Y) \tag{6.29}$$

where $E(Rp)$ and variance (Rp) represent the expected rates of return and variance, respectively. In addition, the summation of weights is assumed to be one $(W_1 + W_2 = 1)$. This assumption is used to guarantee that all available money has been invested in either stock A or stock B.

If John has invested 40 % and 60 % of his portfolio in stock A and stock B, respectively, then the expected rate of return and variance of his portfolio can be calculated in accordance with Eqs. 6.28′ and 6.29′ as

$$E(Rp) = (.6)(.05) + (.4)(-.01) = .026$$

$$\mathrm{Var}(Rp) = (.6)^2(.00625) + (.4)^2(.008) + 2(.6)(.4)(-.00375)$$
$$= .00173$$

These statistics suggest several things:

1. The average rate of return for stock A is higher than that for stock B, and the variance of rates of return for stock A is smaller than that for stock B.
2. The rates of return for stock A are negatively correlated with those of stock B.
3. $E(W_1X + W_2Y) = .026$ represents the average rate of return for a portfolio wherein the different percentage of the money is invested in stock A and in stock B.
4. $\mathrm{Var}(W_1X + W_2Y) = .00173$ represents the variance of a portfolio:

6.9.3 *Expected Value and Variance of Sums of Random Variables*

For n random variables, $X_1, X_2, \ldots, X_n$, Eq. 6.28 can be generalized as

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) \tag{6.30}$$

Thus the expected value of a sum of n random variables is equal to the sum of the expected values of these random variables.

A somewhat analogous relationship exists for variances of *uncorrelated* random variables.[9] If $X_1, X_2, \ldots, X_n$ are n uncorrelated variables, then

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \tag{6.31}$$

[9] Independent variables imply uncorrelated variables, so Eq. 6.31 also holds for independent random variables. Applications of Eqs. 6.30 and 6.31 will be discussed in Appendix 1 and Sect. 21.8 as well as in Appendix 3 of Chap. 21.

Otherwise covariances are needed, as in Eq. 6.29.

Example 6.22 Rates of Return, Variance, and Covariance for JNJ, MRK, and S&P 500. Using annual rates of return for Johnson & Johnson, Merck, and S&P 500 during the period 1970–2009, we calculate the average rate of return, variance,

Data Display

JNJ

-0.681455	0.735557	0.329385	-0.132041	-0.276278	0.120214
-0.119259	0.001873	-0.017184	0.101555	0.286313	-0.619442
0.362091	-0.155059	-0.087818	0.491250	0.272454	0.165022
0.162139	-0.289582	0.230108	0.616842	-0.551293	-0.091578
0.244915	0.584558	-0.409758	0.340804	0.287688	0.124207
0.139719	-0.431191	-0.078010	-0.021172	0.248595	-0.032496
0.122480	0.034602	-0.076435	0.108473		

Data Display

MRK

-0.105661	0.274854	-0.272215	-0.080191	-0.160629	0.064286
0.004346	-0.162669	0.249828	0.097778	0.205990	0.031377
0.031582	0.101627	0.074488	0.493088	-0.080899	0.300894
-0.627023	0.371471	0.185441	0.878595	-0.733803	-0.182990
0.142442	0.753915	0.235280	0.352548	0.409696	-0.537078
0.411867	-0.357447	-0.013313	-0.158294	-0.271964	0.036942
0.418327	0.367425	-0.450781	0.254065		

Data Display

S&P

-0.149428	0.181086	0.110998	-0.016209	-0.228800	0.039952
0.183960	-0.037349	-0.022200	0.072797	0.153092	0.078043
-0.065131	0.339988	0.000312	0.164402	0.264933	0.213633
-0.073354	0.214643	0.036396	0.124301	0.105162	0.085799
0.019960	0.176578	0.237724	0.302655	0.242801	0.222782
0.075256	-0.163282	-0.167680	-0.028885	0.171379	0.067731
0.085510	0.127230	-0.174081	-0.222935		

Fig. 6.11 MINITAB output for Example 6.22

Mean of JNJ

Mean of JNJ = 0.0510197

Standard Deviation of JNJ

Standard deviation of JNJ = 0.322755

Mean of MRK

Mean of MRK = 0.0638299

Standard Deviation of MRK

Standard deviation of MRK = 0.345036

Mean of S&P

Mean of S&P = 0.0687443

Standard Deviation of S&P

Standard deviation of S&P = 0.146719

Covariances: JNJ, MRK, S&P

	JNJ	MRK	S&P
JNJ	0.1041709		
MRK	0.0523250	0.1190496	
S&P	0.0151314	0.0175118	0.0215263

Correlations: JNJ, MRK

Pearson correlation of JNJ and MRK = 0.470

P-Value = 0.002

Fig. 6.11 (continued)

covariance, and correlation coefficient by using MINITAB. The MINITAB outputs are presented in Fig. 6.11.

If we let X, Y, and Z represent rates of return for JNJ, MRK, and S&P 500, respectively, then we find from Fig. 6.11 that $\bar{X} = .0510$, $\bar{Y} = .0638$, $\bar{X} + \bar{Y} = .11$ 48, $S_X^2 = .1042$. $S_Y^2 = .1190$, $S_{XY} = .0523$, and $\rho_{XY} = .470$. Means of these annual rates of return can be used to measure the profitability of the investments; variances

of these annual rates of return can be used as a measure of the risk, or uncertainty, involved in the different investments.

6.10 Summary

In this chapter, we discussed basic concepts and properties of probability distributions for discrete random variables. Important discrete distributions such as the binomial, hypergeometric, and Poisson distributions are discussed in detail. Applications of these distributions in business decisions are also examined.

Using the probability distribution for a random variable, we can calculate the probabilities of specific sample observations. If the probabilities are difficult to calculate, then the means and standard deviations can be used as numerical descriptive measures that enable us to visualize the probability distributions and thereby to make some approximate probability statements about sample observations.

In this chapter, we also discussed joint probability of two random variables. The covariance and coefficient of correlation were presented as means of measuring the degree of relationship between two random variables X and Y.

Questions and Problems

1. A team of students participates in a project. The results show that all students are able to finish the project in 7 days. The distribution for the finishing time is given in the following table.

Finishing time, hours	1	2	3	4	5	6	7
Students	21	43	23	48	31	29	35

   Define x as the finishing time.

   (a) Obtain $P(X = 1)$, $P(X = 2)$, …, $P(X = 7)$.
   (b) Draw the probability distribution.
   (c) Calculate the cumulative function $F_x(x)$.
   (d) Draw the cumulative function.

2. An investment banker estimates the following probability distribution for the earnings per share (EPS) of a firm.

x(EPS)	2.25	2.50	2.75	3.00	3.25	3.50	3.75
$P(x)$	.05	.10	.20	.35	.15	.10	.05

   Calculate the expected value of the EPS.

3. The following table gives the number of unpainted machines in a container. Calculate the mean and standard deviation.

Unpainted	0	1	2	3	4	5	6	7
Probability	.05	.09	.15	.30	.25	.10	.05	.01

4. The following table gives the probability distribution for two random variables: x, which measures the total number of times a person will be ill during a year, and y, the sex of this person, where 1 represents male and 0 female.

x	1	2	3	4	Total
Y					
0	.10	.11	.11	.17	.49
1	.13	.16	.07	.15	.51
	.23	.27	.18	.32	1.00

   (a) Calculate the expected value and standard deviation of the number of times a person will be ill during a year, given that the sex of the person is male. *Hint:* The conditional expectation and conditional standard deviation, which we have not addressed, can be defined as follows:

$$F(X_k) = P(X \leq X_k | Y = 0)$$

$$F(X_k) = P(X \leq X_k | Y = 1)$$

$$\mu_0 = \sum X_k P(X = X_k | Y = 0)$$

$$\mu_1 = \sum X_k, P(X = X_k | Y = 1)$$

$$\sigma_0 = \left[\sum_{i=1}^{n} (X_i - \mu_0)^2 P(X_i | Y = 0) \right]^{1/2}$$

$$\sigma_1 = \left[\sum_{i=1}^{n} (X_i - \mu_1)^2 P(X_i | Y = 1) \right]^{1/2}$$

   (b) Calculate the mean and standard deviation of the number of times a person will be ill during a year, given that the sex of the person is female.

5. The rate of defective items in a production process is 15 %. Assume a random sample of 10 items is drawn from the process. Find the probability that two of them are "defective". Calculate the expected value and variance.
6. Find the mean and standard deviation of the number of successes in binomial distributions characterized as follows:

(a) $n = 20, p = .5$
(b) $n = 100, p = .09$
(c) $n = 30, p = .7$
(d) $n = 50, p = .4$

7. A fair die is rolled 10 times. An "ace" means to roll a "6." Find the probability of getting exactly four aces, of getting five aces, of getting six aces, and of getting four aces or more.
8. A fair coin is tossed eight times.

   (a) Use MINITAB to construct a probability function table.
   (b) What is the probability that you will have exactly four heads?

9. Consider a group of 12 employees of whom five are in management and seven do clerical work. Select at random a sample of size 4. What is the probability that there will be one manager in this sample?
10. A survey was conducted. Of 20 questionnaires that were sent, 12 were completed and returned. We know that 8 of the 20 questionnaires were sent to students and 12 to nonstudents. Only two of the returned questionnaires were from students.

    (a) What is the response rate for each group?
    (b) Assume we have a response in hand. What is the probability that it comes from a student?

11. The number of people arriving at a bank teller's window is Poisson distributed with a mean rate of .75 persons per minute. What is the probability that two or fewer people will arrive in the next 6 min?
12. The Wicker company has one repair specialist who services 200 machines in the shop and repairs machines that break down. The average breakdown rate is $\lambda = .5$ machine per day (or 1 breakdown every 2 days). This technician can fix two machines a day.

    (a) Use MINITAB to construct a probability function, including breakdown frequency from 0 to 10 cases per day. Assume a Poisson distribution.
    (b) Wicker is interested in determining the probability that there will be more than two breakdowns in a day.

13. Two teams are playing each other in seven basketball games. Team A is considered to have a 60–40 % edge over team B. What is the probability that team A will win four or more games?
14. A baseball player usually has four at bats each game. Suppose the baseball player is a lifetime 0.25 hitter. Find the probability that this player will have:

    (a) Two hits out of four at bats
    (b) No hits out of four at bats
    (c) At least one hit out of four at bats

15. A certain insurance salesman sees an average of five customers in a week. Each time he speaks to a customer, he has a 30 % chance of making a deal. What is the probability that he makes five deals after speaking with five customers in a week?
16. A student takes an exam that consists of 10 multiple-choice questions. Each question has five possible answers. Suppose the student knows nothing about the subject and just guesses the answer on each question. What is the probability that this student will answer four out of the 10 questions correctly?
17. A hospital has three doctors working on the night shift. These doctors can handle only three emergency cases in a time period of 30 min. On average, $\frac{1}{2}$ an emergency case arises in each 30-min period. What is the probability that four emergency cases will arise in a 30-min period?
18. An average of three small businesses go bankrupt each month. What is the probability that five small businesses will go bankrupt in a certain month?
19. During each hour, 0.1 % of the total production of paperclips is defective. For a random sample of 500 pieces of the product, what is the chance of finding more than one defective item?
20. The local bank manager has found that one out of every 400 bank loans end up in default. Last year the bank made 400 loans. What is the probability that two bank loans will end up in default?
21. Every week a truckload of springs is delivered to the warehouse you supervise. Every time the springs arrive, you have to measure the strength of 400 springs. You accept the shipment only when there are fewer than 20 bad springs. One day a truckload of springs arrives that contains 10 % bad springs. What is the probability that you will accept the shipment? (Just set up the question. Do not try to solve it.)
22. Returning to question 21, say (1), your company's policy is to accept the shipment only when fewer than two springs (out of 400 springs examined) are bad, and (2) the proportion of the bad springs in the truck is only 0.0001. Under these conditions, what is the probability that you will accept the shipment?
23. Despite your discomfort with statistics, you find yourself employed by a dog food manufacturer to do statistical research for quality control purposes. Your job is to weigh the dog food to determine whether the cans contain the 16 oz of dog food that the label will claim they contain. You pick 25 cans from each hour's production and weigh them. If there are more than two cans that contain less than 16 oz, you are to discard the production from that hour. If in a certain hour, 5 % of the cans of dog food produced actually contain less than 16 oz, what is the probability that the whole hour's production will be discarded?
24. A medical report shows that 5 % of stock brokers suffer stress and need medical attention. There are 10 brokers working for your brokerage house. What is the probability that three of them will need medical attention as a result of stress?
25. There are 38 numbers in the game of roulette. They are 00, 0, 1, 2, . . ., 36. Each number has an equal chance of being selected. In the game, the winning number

is found by a spin of the wheel. Say a gambler bets \$1 on the number 35 three times.

(a) What is the probability that the gambler will win the second bet?
(b) What is the probability that the gambler will win two of the three bets?

26. In the game of roulette, a gambler who wins the bet receives \$36 for every dollar she or he bet. A gambler who does not win receives nothing. If the gambler bets \$1, what is the expected value of the game?
27. A company found that on average, on a given day, .5 % of its employees call in sick. Assume a Poisson distribution. What is the probability that fewer than two of 300 employees will call in sick?
28. Billings Company is considering leasing a computer for the next 3 years. Two computers are available. The net present value of leasing each computer in the next 3 years, under different business conditions, is summarized in the following table.

	Business is	
	Good	Bad
Plan A: big computer	200,000	20,000
Plan B: small computer	150,000	100,000

A consulting company estimates that the chances of having good and of having bad business are 20 % and 80 %, respectively. Compute the expected net present value of leasing a big computer. Compute the expected value of leasing a small computer. What are the variances of these two plans?

29. The makers of two kinds of cola are having a contest in the local shopping mall. Assume that 60 % of the people in this region prefer brand A and 40 % prefer brand B. Ten local residents were randomly selected to test the colas. What is the probability that five of these 10 testers will prefer brand A?
30. In a certain statistics course, the misguided professor is very lenient. He fails about 1 % of the students in the class. Assume that the probability of failing the course follows a Poisson distribution. In a certain year, the professor teaches 400 students. What is the probability that no one fails the course? What is the average number of failing students?
31. A factory examines its work injury history and discovers that the chance of there being an accident on a given workday follows a Poisson distribution. The average number of injuries per workday is .01. What is the probability that there will be three work injuries in a given month (30 days)?
32. The state highway bureau found that during the rush hour, in the treacherous section of a highway, an average of three accidents occur. The probability of there being an accident follows a Poisson distribution. What is the probability that there is no accident in a given day?
33. In question 32, what is the probability that there are no traffic accidents in all five workdays of a week?

34. Of seven prominent financial analysts who are attending a meeting, three are pessimistic about the future of the stock market, and four are optimistic. A newspaper reporter interviews two of the seven analysts. What is the probability that one of these interviewees takes an optimistic view and the other a pessimistic view?
35. After assembly, a finished TV is left turned on for one full day (24 h) to determine whether the product is reliable. On average, two TVs break down each day. Yesterday 500 TVs were produced. What is the probability that less than one TV broke down?
36. A soft drink company argues that its new cola is the favorite soft drink of the next generation. Ten teenagers were picked to test-drink the cola one by one. Assume that five of them liked the new cola and the rest did not.

    (a) What is the probability that the first test-drinker liked the new cola?
    (b) What is the probability that the second test-drinker liked the new cola?
    (c) What is the probability that after five test-drinks, the new product received three yes votes and two no votes?

37. Suppose school records reveal that historically, 10 % of the students in Milton High School have dropped out of school. What is the probability that more than two students in a class of 30 will drop out?
38. An insurance company found that one of 5,000 50-year-old, nonsmoking males will suffer a heart attack in a given year. The company has 50,000 50-year-old, nonsmoking male policyholders. What is the probability that fewer than three such policyholders will suffer a heart attack this year?
39. Suppose that of 40 salespersons in a company, 10 are females and the rest males. Five of them are randomly chosen to attend a seminar. What is the probability that three females and two males are chosen?
40. Consider a single toss of a fair coin, and define X as the number of heads that come up on that toss. Then X can be 0 or 1, with a probability of 50 %. The expected value of X is $\frac{1}{2}$. Can we "expect" to get $\frac{1}{2}$ a head when we toss the coin? If not, how should we interpret the concept of the expected value?
41. What is a random variable? What is a discrete random variable? What is a continuous random variable? Give some examples of discrete random variables.
42. Tell whether each of the following is a discrete or a continuous random variable:

    (a) The number of beers sold at a bar during a particular week
    (b) The length of time it takes a person to drive 50 miles
    (c) The interest rate on 3-month Treasury bills
    (d) The number of products returned to a store on a particular day

43. An analyst calculates the probability of McGregor stock going up in value for any month as .6 and the probability of the same stock going down in any month

as .4. Calculate the probability that the stock will go up in value in exactly 7 months during a year. (Assume independence.)

44. Using the information from question 43, compute the probability that the stock will go up in at least 7 months during the year.
45. What is a Bernoulli trial? Give some examples of a Bernoulli trial related to the binomial distribution.
46. What is the Poisson distribution? Give some examples of situations wherein it would be appropriate to use the Poisson distribution. Compare the Poisson approximation to the binomial distribution.
47. Suppose Y represents the number of times a homemaker stops by the local convenience store in a week. The probability distribution of Y follows. Find the expected value and variance of Y.

y	Probability
0	.15
1	.25
2	.25
3	.20
4	.15

48. The managers of a grocery store are interested in knowing how many people will shop in their store in a given hour. Suppose they collect data and find that the average number of people who enter the store in any 15-min period is 12. Find the probability that eight people will enter the store in any 15-min period. What is the probability that no more than eight people will enter the store in any 15-min period?
49. Suppose you are tossing a fair coin 20 times. What is the probability that you will toss exactly five heads? What is the probability that you will toss five or fewer heads?
50. Calculate the mean and variance for the distribution given in question 49.
51. You are rolling a six-sided fair die eight times. What is the probability that you will roll exactly two sixes? What is the probability that you will roll two or fewer sixes?
52. Calculate the mean and variance for the distribution given in question 51.
53. Doctors at the Centers for Disease Control estimate that 30 % of the population will catch the Tibetan flu. What is the probability that in a sample of 10 people, exactly three will catch the flu? What is the probability that three or fewer people in this sample will catch the flu? (Assume that the conditions of a Bernoulli process apply.)
54. A golfer enters a long-driving contest in which he wins if he drives the golf ball 300 yards or more and loses if he drives it less than 300 yards. Assume that every time he hits a golf ball, he has a 40 % chance of driving it over 300 yards.

If the golfer gets to hit four balls and needs only one 300-yard drive to win, what is the probability that he will win?

55. A phone marketing company knows that the number of people who answer the phone between 10:00 and 10:15 a.m. has a Poisson distribution. The average number is eight. What is the probability that the phone company will reach exactly 10 people when it calls during this period? What is the probability that it will reach exactly three people?
56. A market survey shows that 75 % of all households own a VCR. Suppose 100 households are surveyed.
    (a) What is the probability that none of the households surveyed owns a VCR?
    (b) What is the probability that exactly 75 of the households surveyed own a VCR?
    (c) Suppose X is the number of households that own a VCR. Compute the mean and variance for X.
57. You are given the following information about a stock:

S = \$100	Price of stock
X = \$10 l	Exercise price for a call option on the stock
$r = .005$	Interest rate per month
$n = 5$	Number of months until the option expires
$u = 1.10$	Amount of increase if stock goes up
$d = .95$	Amount of decrease if stock goes down

    Calculate the value of the call option if the stock goes up in 3 out of the 5 months.
58. Answer question 57 when $u = 1.20$. How does a change in the amount of increase if the stock goes up affect the value of the call option?
59. Answer question 57 when $d = .85$. How does a change in the amount of decrease if the stock goes down affect the value of the call option?
60. Answer question 57 when $X = \$95$. How does a change in the exercise price affect the value of the call option?
61. Answer question 57 when $S = \$110$. How does a change in the value of the stock affect the value of the call option?
62. Answer question 57 again, finding the value of the option if the stock goes up in 4 out of the 5 months.
63. Suppose a box is filled with 25 white balls and 32 red halts. Find the probability of drawing six red balls and four white balls in 10 draws without replacement.
64. Redo question 63 *with* replacement.
65. You are drawing five cards from a standard deck of cards with replacement. You win if you draw at least three red cards in five draws. What is the probability of your winning?
66. Two tennis players are playing in a final-set tie breaker. Player A is considered to have a 70–30 % edge over player B. The player who wins seven of 13 points

will win the championship. What is the probability that player B will win the championship?

67. A car sales representative sees an average of four customers in a day. Each time she talks to a customer, she has a 25 % chance of making a deal. What is the chance that she will make four deals after talking to four customers in a day?
68. Again consider the car sales rep in question 67. What is the probability that she will make at least two sales after speaking to four customers?
69. Again consider the car sales rep in question 67. What is the probability that she will make exactly two sales after speaking to four customers?
70. The following test is given to people who claim to have extrasensory perception (ESP). Five cards with different shapes on them are hidden from the person. A card is randomly drawn, and the person is then supposed to guess (or use ESP to determine) the shape on the card. Suppose that this test is administered 10 times with replacement. What is the probability that the subject will get five correct? Do you think getting more than five out of 10 correct supports the subject's claim to be endowed with ESP?
71. Again consider the test in question 70. What is the probability that a person taking the test will get eight out of 20 correct? Do you believe that a person who gets eight out of 20 correct has ESP?
72. Say we toss two six-sided dice and let the random variable be the total number of dots observed.

    (a) Calculate both the probability and the cumulative probability distributions.
    (b) Draw a graph associated with probability distribution you obtained in part (a).

73. (a) Use the monthly rates of return for both GM and Ford listed in Fig. 6.11 to calculate the correlation coefficient between the monthly rates of return for these two companies.
    (b) Using the results you obtained in part (a) and some other statistics listed in Fig. 6.11, discuss how these two securities are related and how this information can be used to make investment decisions.

74. The following table exhibits the monthly rate of return of S&P 500 and American Express. Use the MINITAB program to:

    (a) Calculate the mean and standard deviation of both returns.
    (b) Calculate the correlation coefficient between these two sets of monthly returns.
    (c) Explain the results you get from the two questions above.

		S&P 500	AMEX
1989	9	−.65	−2.69
	10	−2.52	12.73
	11	16.54	−2.97

(continued)

(continued)

		S&P 500	AMEX
	12	21.42	−1.11
1990	1	−6.88	−15.25
	2	8.54	−2.10
	3	24.26	−11.59
	4	−2.69	6.23
	5	9.20	6.93
	6	−.89	5.91

75. A production process produces 0.05 % defective parts. A sample of 10,000 parts from the production is selected. In (a) and (b), what is the probability that:

    (a) The sample contains exactly two defective parts?
    (b) The sample contains no defective parts?
    (c) Find the expected number of defective parts.
    (d) Find the standard deviation for the number of defective parts.

76. The results of a survey of married couples and the number of children they had are shown below.

Number of children	Probability
0	0.150
1	0.125
2	0.500
3	0.175
4	0.050

    Determine the expected number of children and the standard deviation for the number of children.

77. The average number of calls received by an operator in a 30-min period is 12.

    (a) What is the probability that between 17:00 and 17:30 the operator will receive exactly eight calls?
    (b) What is the probability that between 17:00 and 17:30 the operator will receive more than nine calls but fewer than 15 calls?
    (c) What is the probability that between 17:00 and 17:30 the operator will receive no calls?

78. In a lot of 200 parts, 50 of them are defective. Suppose a sample of 10 parts is selected at random, what is the probability that two of them are defective? What is the expected number of defective parts? What is the standard deviation of the number of defective parts?

Appendix 1: The Mean and Variance of the Binomial Distribution

Let X_i represent the Bernoulli random variable; then the random variables X_1, X_2, . . ., X_n are independent Bernoulli variables. From Eqs. 6.3 and 6.4, we know that the mean of a Bernoulli variable is

$$E(X_i) = \sum_{i=1}^{2} x_i P(x_i) = 0(1-p) + 1(p) = p \tag{6.32}$$

and that the variance is

$$\begin{aligned} \mathrm{Var}(X_i) &= \sum_{i=1}^{2} (x_i - \mu_X)^2 P(x_i) \\ &= (0-p)^2(1-p) + (1-p)^2 p = p(1-p) \end{aligned} \tag{6.33}$$

To find the mean and variance of the binomial distribution, we use the fact that the binomial random variable can be expressed as the sum of independent Bernoulli random variables (X_i):

$$X = X_1 + X_2 + \cdots + X_n \tag{6.34}$$

From Eqs. 6.30 and 6.32, we can derive Eq. 6.11 as

$$\begin{aligned} E(X) = \mu &= E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= np \end{aligned} \tag{6.35}$$

Because the X_i variables are statistically independent of one another, the variance of their sum is equal to the sum of their variances. Therefore, following Eq. 6.31, we can derive Eq. 6.12 as

$$\sigma^2 = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) = np(1-p)$$

Appendix 2: Applications of the Binomial Distribution to Evaluate Call Options

In this appendix, we show how the binomial distribution is combined with some basic finance concepts to generate a model for determining the price of stock options.

What Is an Option?

In the most basic sense, an *option* is a contract conveying the right to buy or sell a designated security at a stipulated price. The contract normally expires at a predetermined date. The most important aspect of an option contract is that the purchaser is under no obligation to buy; it is, indeed, an "option." This attribute of an option contract distinguishes it from other financial contracts. For instance, whereas the holder of an option may let his or her claim expire unused if he or she so desires, other financial contracts (such as future and forward contracts) obligate their parties to fulfill certain conditions.

A *cad option* gives its owner the right to buy the underlying security a *put option* the right to sell. The price at which the stock can be bought (for a call option) or sold (for a put option) is known as the exercise price.

The Simple Binomial Option Pricing Model

Before discussing the binomial option pricing model, we must recognize its two major underlying assumptions. First, the binomial approach assumes that trading takes place in discrete time – that is, on a period-by-period basis. Second, it is assumed that the stock price (the price of the underlying asset) can take on only two possible values each period; it can go up or go down.

Say we have a stock whose current price per share S can advance or decline during the next period by a factor of either u (up) or d (down). This price either will increase by the proportion $u - 1 \geq 0$ or will decrease by the proportion $1 - d$, $0 < d < 1$. Therefore, the value S in the next period will be either uS or dS. Next, suppose that a call option exists on this stock with a current price per share of C and an exercise price per share of X and that the option has one period left to maturity. This option's value at expiration is determined by the price of uS underlying stock and the exercise price X. The value is either

$$C_u = Max(0, uS - X) \tag{6.36}$$

or

$$C_d = Max(0, dS - X) \tag{6.37}$$

Why is the call worth $\text{Max}(0, uS - X)$ if the stock price is uS? The option holder is not obliged to purchase the stock at the exercise price of X, so she or he will exercise the option only when it is beneficial to do so. This means the option can never have a negative value. When is it beneficial for the option holder to exercise the option? When the price per share of the stock is greater than the price per share at which he or she can purchase the stock by using the option, which is the exercise price X? Thus, if the stock price uS exceeds the exercise price X, the investor can

exercise the option and buy the stock. Then he or she can immediately sell it for uS, making a profit of $uS - X$ (ignoring commission). Likewise, if the stock price declines to dS, the call is worth $\text{Max}(0, dS - X)$.

Also for the moment, we will assume that the risk-free interest rate for both borrowing and lending is equal to r percent over the one time period and that the exercise price of the option is equal to X.

To intuitively grasp the underlying concept of option pricing, we must set up a *risk-free portfolio* – a combination of assets that produces the same return in every state of the world over our chosen investment horizon. The investment horizon is assumed to be one period (the duration of this period can be any length of time, such as an hour, a day, and a week). To do this, we buy h shares of the stock and sell the call option at its current price of C.[10] Moreover, we choose the value of h such that our portfolio will yield the same payoff whether the stock goes up or down:

$$h(uS) - C_u = h(dS) - C_d \tag{6.38}$$

By solving for h, we can obtain the number of shares of stock we should buy for each call option we sell:

$$h = \frac{C_u - C_d}{(u - d)S} \tag{6.39}$$

Here h is called the *hedge ratio*. Because our portfolio yields the same return under either of the two possible states for the stock, it is without risk and therefore should yield the risk-free rate of return, r percent, which is equal to the risk-free borrowing and lending rate. The condition must be true; otherwise, it would be possible to earn a risk-free profit without using any money. Therefore, the ending portfolio value must be equal to $(1 + r)$ times the beginning portfolio value, $hS - C$:

$$(1 + r)(hS - C) = h(uS) - C_u = h(dS) - C_d \tag{6.40}$$

Note that S and C represent the beginning values of the stock price and the option price, respectively.

Setting $R = 1 + r$, rearranging to solve for C, and using the value of h from Eq. 6.39, we get

$$C = \left[\left(\frac{R - d}{u - d}\right)C_u + \left(\frac{u - R}{u - d}\right)C_d\right] \Big/ R. \tag{6.41}$$

where $d < r < u$. To simplify this equation, we set

[10] To sell the call option means to write the call option. If a person writes a call option on stock A, then he or she is obliged to sell at exercise price X during the contract period.

Table 6.13 Possible option values at maturity

Today		
Stock (S)	*Option (C)*	Next period (maturity)
$100	C	$uS = \$110\ C_u = \text{Max}\ (0, uS - X)$ $= \text{Max}(0, 110 - 100)$ $= \text{Max}(0, 10)$ $= \$10$ $dS = \$90\ C_d = \text{Max}\ (0, dS - X)$ $= \text{Max}\ (0, 90 - 100)$ $= \text{Max}(0, -10)$ $= \$0$

$$p = \frac{R - d}{u - d} \quad \text{so} \quad 1 - p = \left\{\frac{u - R}{u - d}\right\} \tag{6.42}$$

Thus we get the option's value with one period to expiration:

$$C = [pC_u + (1 - p)C_d]/R \tag{6.43}$$

This is the binomial call option valuation formula in its most basic form. In other words, this is the binomial option valuation formula with one period to expiration of the option.

To illustrate the model's qualities, let's plug in the following values while assuming the option has one period to expiration. Let

$$\begin{aligned} X &= \$100 \\ S &= \$100 \\ u &= (1.10), \ \text{so}\ uS = \$110 \\ d &= (.90), \ \ \text{so}\ dS = \$90 \\ R &= 1 + r = 1 + .07 = 1.07 \end{aligned}$$

First we need to determine the two possible option values at maturity, as indicated in Table 6.13.

Next we calculate the value of p as indicated in Eq. 6.42:

$$p = \frac{1.07 - .90}{1.10 - .90} = .85 \ \text{so}\ 1 - p = \frac{1.10 - 1.07}{1.10 - .90} = .15$$

Solving the binomial valuation equation as indicated in Eq. 6.43, we get

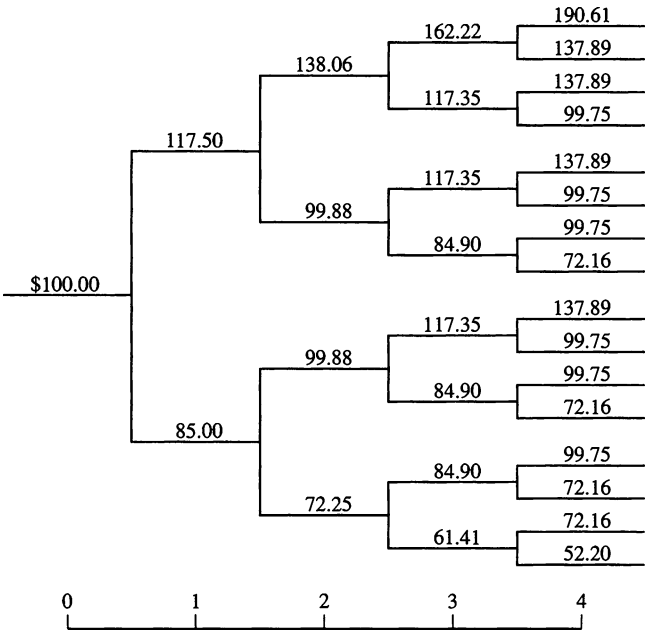


Fig. 6.12 Price path of underlying stock (Source: R. J. Rendelman, Jr., and B. J. Bartter(1979). "Two-State Option Pricing," *Journal of Finance* 34 (December), 1096)

$$C = [.85(10) + .15(0)]/1.07$$
$$= \$7.94$$

The correct value for this particular call option today, under the specified conditions, is \$7.94. If the call option does not sell for \$7.94, it will be possible to earn arbitrage profits. That is, it will be possible for the investor to earn a risk-free profit while using none of his or her own money. Clearly, this type of opportunity cannot continue to exist indefinitely.

[11] This section is essentially based on Cheng F. Lee, Joseph E. Finnerty, and Donald H. Wort (1990) *Security Analysis and Portfolio Management* (Glenview. III.: Scott. Foresman), Chapter 15. Copyright © 1990 by Cheng F. Lee. Joseph E. Finnerty, and Donald H. Wort. Reprinted by permission of Harper Collins Publishers.

The Generalized Binomial Option Pricing Model

Suppose we are interested in the case where there is more than one period until the option expires.[11] We can extend the one-period binomial model to consideration of two or more periods. Because we are assuming that the stock follows a binomial process, from one period to the next, it can only go up by a factor of u or go down by a factor of d. After one period the stock's price is either uS or dS. Between the first and second periods, the stock's price can once again go up by u or down by d, so the possible prices for the stock two periods from now are uuS, udS, and ddS. This process is demonstrated in tree diagram form (Fig. 6.12) in Example 6.23 later in this appendix.

Note that the option's price at expiration, two periods from now, is a function of the same relationship that determined its expiration price in the one-period model. More specifically, the call option's maturity value is always

$$C_T = \text{Max}[0, S_T - X] \tag{6.44}$$

where T designates the maturity date of the option.

To derive the option's price with two periods to go ($T = 2$), it is helpful as an intermediate step to derive the value of C_u and C_d with one period to expiration when the stock price is uS and dS, respectively:

$$C_u = [pC_{uu} + (1-p)C_{ud}]/R \tag{6.45}$$

$$C_d = [pC_{du} + (1-p)C_{dd}]/R \tag{6.46}$$

Equation 6.45 tells us that if the value of the option after one period is C_u, the option will be worth either C_{uu} (if the stock price goes up) or C_{ud} (if stock price goes down) after one more period (at its expiration date). Similarly, Eq. 6.46 shows that if the value of the option is C_d after one period, the option will be worth either C_{du} or C_{dd} at the end of the second period. Replacing C_u and C_d in Eq. 6.43 with their expressions in Eqs. 6.45 and 6.46, respectively, we can simplify the resulting equation to yield the two-period equivalent of the one-period binomial pricing formula, which is

$$C = \left[p^2 C_{uu} + 2p(1-p)C_{ud} + (1-p)^2 C_{dd}\right]/R^2 \tag{6.47}$$

In Eq. 6.47, we used the fact that $C_{ud} = C_{du}$ because the price will be the same in either case.

We know the values of the parameters S and X. If we assume that R, u, and d will remain constant over time, the possible maturity values for the option can be determined exactly. Thus, deriving the option's fair value with two periods to maturity is a relatively simple process of working backward from the possible maturity values.

Using this same procedure of going from a one-period model to a two-period model, we can extend the binomial approach to its more generalized form, with n periods to maturity:

$$C = \frac{1}{R^n} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \text{Max}\left[0, u^k d^{n-k} S - X\right] \tag{6.48}$$

To actually get this form of the binomial model, we could extend the two-period model to three periods, then from three periods to four periods, and so on. Equation 6.48 would be the result of these efforts. To show how Eq. 6.48 can be used to assess a call option's value, we modify the example as follows: $S = \$100$, $X = \$100$, $R = 1.07$, $n = 3$, $u = 1.1$, and $d = .90$.

First we calculate the value of p from Eq. 6.42 as .85, so $1 - p$ is. 15. Next we calculate the four possible ending values for the call option after three periods in terms of Max[0, $u^k d^{n-k} S - X$]:

$$C_1 = \text{Max}\left[0, (1.1)^3 (.90)^0 (100) - 100\right] = 33.10$$
$$C_2 = \text{Max}\left[0, (1.1)^2 (.90)(100) - 100\right] = 8.90$$
$$C_3 = \text{Max}\left[0, (1.1)(.90)^2 (100) - 100\right] = 0$$
$$C_4 = \text{Max}\left[0, (1.1)^0 (.90)^3 (100) - 100\right] = 0$$

Now we insert these numbers (C_1, C_2, C_3, and C_4) into the model and sum the terms:

$$\begin{aligned} C &= \frac{1}{(1.07)^3} \left[\frac{3!}{0!3!} (.85)^0 (.15)^3 \times 0 + \frac{3!}{1!2!} (.85)^1 (.15)^2 \times 0 \right. \\ &\quad \left. + \frac{3!}{2!1!} (.85)^2 (.15)^1 \times 8.90 + \frac{3!}{3!0!} (.85)^3 (.15)^0 \times 33.10 \right] \\ &= \frac{1}{1.225} \left[0 + 0 + \frac{3 \times 2 \times 1}{2 \times 1 \times 1} (.7225)(.15)(8.90) \right. \\ &\quad \left. + \frac{3 \times 2 \times 1}{3 \times 2 \times 1 \times 1} \times (.61413)(1)(33.10) \right] \\ &= \frac{1}{1.225} [(.32513 \times 8.90) + (.61413 \times 33.10)] \\ &= \$18.96 \end{aligned}$$

As this example suggests, working out a multiple-period problem by hand with this formula can become laborious as the number of periods increases. Fortunately, programming this model into a computer is not too difficult.

Now let's derive a binomial option pricing model in terms of the cumulative binomial density function. As a first step, we can rewrite Eq. 6.48 as

$$C = S\left[\sum_{k=m}^{n} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \frac{u^k d^{n-k}}{R^n}\right] - \frac{X}{R^n}\left[\sum_{k=m}^{n} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}\right] \tag{6.49}$$

This formula is identical to Eq. 6.48 except that we have removed the Max operator. In order to remove the Max operator, we need to make $u^k d^{n-k} S - X$ positive, which we can do by changing the counter in the summation from $k = 0$ to $k = m$. What is m? It is the minimum number of upward stock movements necessary for the option to terminate "in the money" (i.e., $u^k d^{n-k} S - X > 0$). How can we interpret Eq. 6.49? Consider the second term in brackets; it is just a cumulative binomial distribution with parameters of n and p.[12] Likewise, via a small algebraic manipulation, we can show that the first term in the brackets is also a cumulative binomial distribution. This can be done by defining $p' \equiv (u/R)p$ and $1 - p' \equiv (d/R)(1-p)$.[13] Thus

$$p^k (1-p)^{n-k} \frac{u^k d^{n-k}}{R^n} = p'^k (1-p')^{n-k}$$

Therefore, the first term in brackets is also a cumulative binomial distribution with parameters of n and p'. Using Eq. 6.10 in the text, we can write the binomial call option model as

$$C = SB_1(n, p', m) - \frac{X}{R^n} B_2(n, p, m) \tag{6.50}$$

where

[12] Note that this is not exactly a cumulative binomial distribution as defined by a statistician. Strictly speaking,

$$1 - [\,] = \sum_{k=0}^{m-1} \frac{n!}{k!(n-k)!} p^k (1-p)^{m-k}$$

is a cumulative binomial distribution.

[13] Because $u < R < d$,

$$(u/R)P = \frac{1 - d/R}{1 - d/u} < 1$$

$$B_1(n, p', m) = \sum_{k=m}^{n} {}_nC_k p'^k (1-p')^{n-k}$$

$$B_2(n, p, m) = \sum_{k=m}^{n} {}_nC_k p^k (1-p)^{n-k}$$

and m is the minimum amount of time the stock has to go up for the investor to finish *in the money* (i.e., for the stock price to become larger than the exercise price).

In this appendix, we showed that by employing the definition of a call option and by making some simplifying assumptions, we can use the binomial distribution to find the value of a call option. In the next chapter, we will show how the binomial distribution is related to the normal distribution and how this relationship can be used to derive one of the most famous valuation equations in finance, the Black–Scholes option pricing model.

Example 6.23 A Decision Tree Approach to Analyzing Future Stock Price. By making some simplifying assumptions about how a stock's price can change from one period to the next, it is possible to forecast the future price of the stock by means of a decision tree. To illustrate this point, let's consider the following example.

Suppose the price of company A's stock is currently \$100. Now let's assume that from one period to the next, the stock can go up by 17.5 % or go down by 15 %. In addition, let us assume that there is a 50 % chance that the stock will go up and a 50 % chance that the stock will go down. It is also assumed that the price movement of a stock (or of the stock market) today is completely independent of its movement in the past; in other words, the price will rise or fall today by a random amount. A sequence of these random increases and decreases is known as a *random walk*. Method of testing the randomness of stock rates of return will be discussed in Sect. 17.8 of Chap. 17.

Given this information, we can lay out the paths that the stock's price may take. Figure 6.12 shows the possible stock prices for company A for four periods.

Note that in period 1 there are two possible outcomes: the stock can go up in value by 17.5 % to \$117.50 or down by 15 % to \$85.00. In period 2, there are four possible outcomes. If the stock went up in the first period, it can go up again to \$138.06 or down in the second period to \$99.88. Likewise, if the stock went down in the first period, it can go down again to \$72.25 or up in the second period to \$99.88. Using the same argument, we can trace the path of the stock's price for all four periods.

If we are interested in forecasting the stock's price at the end of period 4, we can find the average price of the stock for the 16 possible outcomes that can occur in period 4:

$$\begin{aligned}\bar{P} &= \frac{\sum_{i=1}^{16} P_i}{16} = \frac{190.61 + 137.89 + \cdots + 52.20}{16} \\ &= \$105.09\end{aligned}$$

We can also find the standard deviation for the stock' return:

$$\begin{aligned}\sigma_P &= \left[\frac{(190.61 - 105.09)^2 + \cdots + (52.20 - 105.09)^2}{16}\right]^{1/2} \\ &= \$34.39\end{aligned}$$

$\bar{p}$ and σ_P can be used to predict the future price of stock A.

Chapter 7
The Normal and Lognormal Distributions

Chapter Outline

Key Terms

Continuous random variable	Cumulative uniform density function
Probability mass function	Uniform distribution
Cumulative probability	Coefficient of variation
Cumulative distribution function	Standard normal distribution
Probability density function	*Z* score
Normal distribution	Lognormal distribution
Normal probability density function	Option pricing model
	Put-call parity

7.1 Introduction

In Chap. 6, we discussed discrete random variables and their distributions. Particularly, we focused on the means and variances of binomial, hypergeometric, and Poisson distributions. Although the distributions derived from these discrete

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_7, © Springer Science+Business Media New York 2013

random variables are useful, they are limited. And therefore, statisticians have derived several important continuous distributions to substitute for and/or complement the discrete distributions. The normal distribution is the first important continuous distribution discussed in this chapter. Examples of continuous random variables include the number of miles a car travels on 1 gal of gas and the exact weight of a box of cereal.

The lognormal distribution, a transformation of the normal distribution, is the second important continuous distribution we will examine. It is useful in many business and economic analyses. Because the lognormal distribution is valid only for nonnegative values of the random variable, it is more appropriate than the normal distribution for describing the distribution of a stock's price.

In this chapter, we also discuss how the normal distribution can be used to approximate both binomial and Poisson distributions when the sample size is large.

7.2 Probability Distributions for Continuous Random Variables

7.2.1 Continuous Random Variables

Unlike the values of discrete random variables, which are limited to a finite or countable number of distinct (integer) values, values of continuous variables are *not* limited to being integers; theoretically, they are infinitely divisible. A *continuous random variable* may take on any value within an interval, as we noted in Sect. 6.2. Measures of height, weight, time, distance, and temperature fit naturally into this category. In general, specific probabilities cannot be assigned to individual values of continuous random variables. The probability that any one specific value will occur for a continuous random variable is zero. For example, the probability that today's temperature is exactly 83.231° is zero, because temperature is regarded as a continuous variable.

One may argue that in the real world, all data are discrete. For example, if a scale permits determination of weight only to the nearest thousandth of a pound, then any resulting data will be discrete in units of thousandths of pounds. Despite the limitations of measuring instruments, however, it is useful in many instances to use continuous mathematical models that treat certain discrete variables as continuous. If we use a continuous mathematical model of heights of individuals, where the underlying data are discrete, we may conceive of this model not as a convenient approximation but rather as a model of reality that is more accurate than the discrete data from which the model was derived. In sum, even though measurement limitations make continuous data discrete, we are going to treat data as continuous because the model that results when we do so is more accurate.

7.2.2 *Probability Distribution Functions for Discrete and Continuous Random Variables*

We shall now consider experiments for which the theoretical set of possible outcomes forms a continuous interval on the real number line. Note that such observations are often rounded off so that the set of observations may seem to come from a finite set of real numbers. For such an experiment, we should consider conceptual sample spaces that are intervals of finite or infinite length. In this section, we contrast probability distribution functions for continuous random variables with those for discrete random variables, which we discussed in Sect. 6.3.

7.2.2.1 Approximation of a Histogram by a Continuous Curve

In Chap. 6, we showed that the probability distribution of a discrete random variable can be represented by a histogram. It can also be shown that a histogram can be approximated by a continuous curve. Now we use a fair coin-tossing example to demonstrate the meaning of a graph of the probability distribution of a continuous random variable. This example demonstrates the relationship between the probability distribution of a discrete variable and the probability distribution of a continuous variable.

Example 7.1 Using a Continuous Curve to Approximate the Histogram of a Fair Coin-Tossing Experiment. The binomial distribution discussed in Chap. 6 is an example of a probability distribution of a discrete random variable. To get better insight into the meaning of a graph of the probability distribution of a discrete versus a continuous random variable, we first graph a binomial distribution as a histogram.

If we toss a fair coin four times, the probabilities of our getting 0, 1,2, 3, and 4 heads, respectively, can be calculated by using the binomial formula, Eq. 6.9:

$$P(X_1 = 4 \text{ tails}) = \binom{4}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4 = \frac{1}{16}$$

$$P(X_2 = 1 \text{ head and } 3 \text{ tails}) = \binom{4}{1}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 = \frac{4}{16}$$

$$P(X_3 = 2 \text{ heads and } 2 \text{ tails}) = \binom{4}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = \frac{6}{16}$$

$$P(X_4 = 3 \text{ heads and } 1 \text{ tail}) = \binom{4}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) = \frac{4}{16}$$

$$P(X_5 = 4 \text{ heads}) = \binom{4}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 = \frac{1}{16}$$

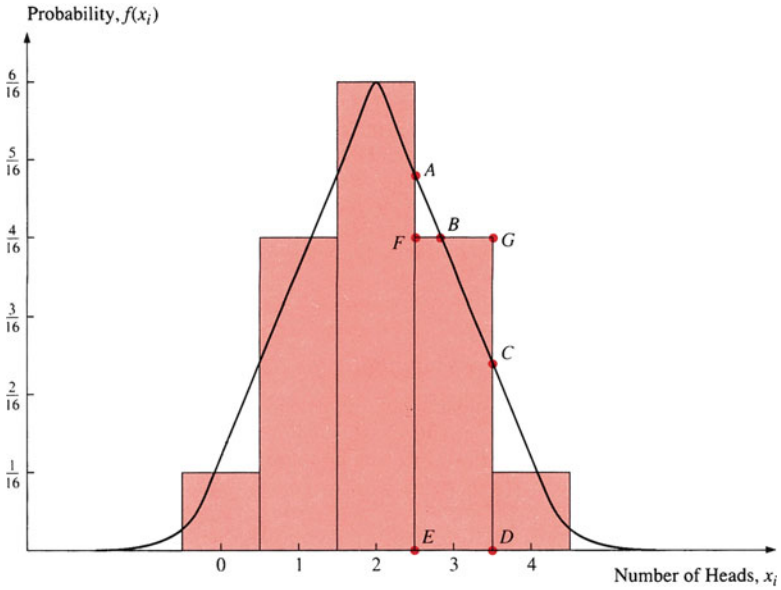


Fig. 7.1 Approximation of a binomial distribution histogram by a continuous curve ($n = 4$, $p = 1/2$)

This is a binomial distribution with $p = \frac{1}{2}$ and $n = 4$. The graph of this histogram is shown in Fig. 7.1. Using this histogram, we interpret 0, 1, 2, 3, and 4 heads not as discrete values but as midpoints of five classes whose respective limits are $-\frac{1}{2}$ to $\frac{1}{2}$, $\frac{1}{2}$ to $1\frac{1}{2}$, $1\frac{1}{2}$ to $2\frac{1}{2}$, $2\frac{1}{2}$ to $3\frac{1}{2}$, and $3\frac{1}{2}$ to $4\frac{1}{2}$. The probabilities or relative frequencies associated with these classes are represented in the graph by the areas of rectangles or bars. Thus, because the rectangle for the class interval $1\frac{1}{2}$ to $2\frac{1}{2}$ has 1.5 times the area of that for the interval $2\frac{1}{2}$ to $3\frac{1}{2}$, it represents 1.5 times the probability. We can draw a continuous curve over the histogram and make the total area of this curve equal to the total area of the sum of five rectangles, which is 1.

The curve would pass through the rectangle at B for 3 heads, as shown in Fig. 7.1. For example, area $ABCDE$ represents the probability of the class with 3 heads in terms of a continuous-variable curve. This is due to the fact that ΔABF is approximately equal to ΔBCG. The area under the curve bounded by the class limits for any given class represents the probability of occurrence of that class.

If n increased (say, to 6 or 200), the width of the rectangles would decrease, and the corresponding shape of the histogram would approach that of a continuous curve more closely. Just as the total area of the rectangles in a histogram, representing a discrete random variable distribution, is equal to 1, so is the total area under the continuous curve representing a continuous random variable distribution.

We will use this example first to review the probability distributions for discrete variables. Then we will develop probability distributions for continuous variables by contrasting them with the probability functions of discrete variables discussed here.

7.2.2.2 Cumulative Probability and Cumulative Distribution Function for Discrete Random Variables

Let the value of the *probability mass function* (PMF) of a discrete random variable X at x be denoted as $P(x)$. In accordance with Eq. 6.2, the *cumulative probability* for X, which is the probability that X will assume a value less than or equal to a given number x_k, can be defined as

$$F(x_k) = P(X \leq x_k) = P(x_1) + P(x_2) + \cdots + P(x_k) = \sum_{i=1}^{k} P(x_i) \tag{6.2}$$

where $x_1 < x_2 \ldots < x_k$.

Now let us see how cumulative probabilities are calculated.

Example 7.2 Cumulative Probability for Fair Coin-Tossing Experiments. In the coin-tossing case discussed in Example 7.1, $P(x_1) = \frac{1}{16}$, $P(x_2) = \frac{4}{16}$, $P(x_3) = \frac{6}{16}$ $P(x_4) = \frac{4}{16}$, and $P(x_5) = \frac{1}{16}$. We calculate the cumulative probabilities $F(x_2)$, $F(x_4)$, and $F(x_5)$ by using Eq. 6.2:

$$F(x_2) = \frac{1}{16} + \frac{4}{16} = \frac{5}{16}$$
$$F(x_4) = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} = \frac{15}{16}$$
$$F(x_5) = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{4}{16} + \frac{1}{16} = 1$$

If the values outside the range of X (i.e., the values smaller than x_1 or larger than x_k) occur with probability equal to zero, then in accordance with Eq. 6.2a, the *cumulative distribution function* (CDF) of X can be written as

$$F(x_k) = \sum_{i=-\infty}^{k} P(x_i) \tag{6.2a}$$

The probability that X lies between a and b is

$$P(a \leq X \leq b) = F(b) - F(a)$$
$$= \sum_{i=1}^{b} P(x_i) - \sum_{i=1}^{a} P(x_i) \tag{7.1}$$

where $F(a)$ and $F(b)$are cumulative probabilities at $X = b$ and $X = a$, respectively.

Example 7.3 Cumulative Distribution Function for the Tossing of a Fair Coin. For the fair coin-tossing case discussed in Example 7.1, the CDFs calculated in accordance with Eq. 6.2 are presented in Table 7.1. Using the probabilities of Table 7.1, we calculate the probability that lies between x_4 and x_2 as

$$P(x_2 \leq X \leq x_4) = F(x_4) - F(x_2) = \frac{15}{16} - \frac{5}{16} = \frac{10}{16}$$

Table 7.1 Cumulative distribution function for coin tossing

Possible values of X	$F(x)$
0	1/16
1	5/16
2	11/16
3	15/16
4	1

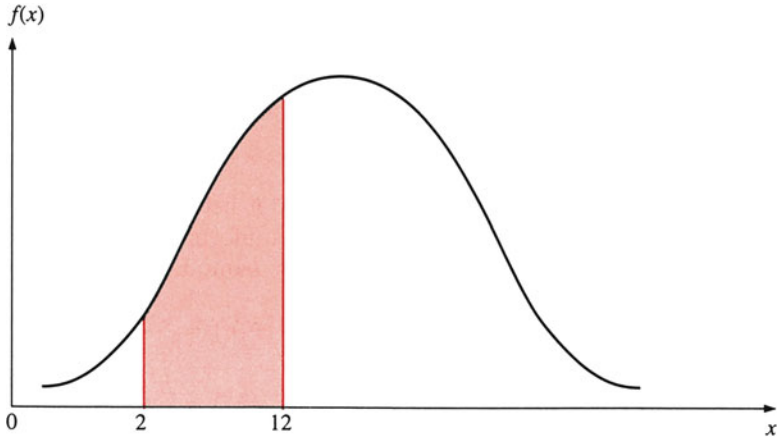


Fig. 7.2 Probability density function

7.2.2.3 Probability Distributions for Continuous Random Variables

The *probability density function* (PDF) for a continuous random variable is a curve, $f(x)$, that shows the probability of a range of values as the area under the curve. For example, the probability of the birth weights of infants being between 2 and 12 lb can be written $P(2 < X < 12)$. Graphically, it is represented by the shaded area of Fig. 7.2.

For continuous random variables, the probability that X has a value between a and b is written $P(a \leq X \leq b)$. This probability is equal to $P(a < X < b)$, because the probability at a point is considered to be zero; that is, $P(X = a) = 0$ and $P(X = b) = 0$. Thus, we can write $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$.

Analogously to Eq. 7.1, we define the probability that X lies between a and b for a continuous random variable as

$$P(a < X < b) = F(b) - F(a) \tag{7.2}$$

To show how Eq. 7.2 can be used to calculate the probability of a continuous variable, we first discuss the simplest continuous cumulative density function, the

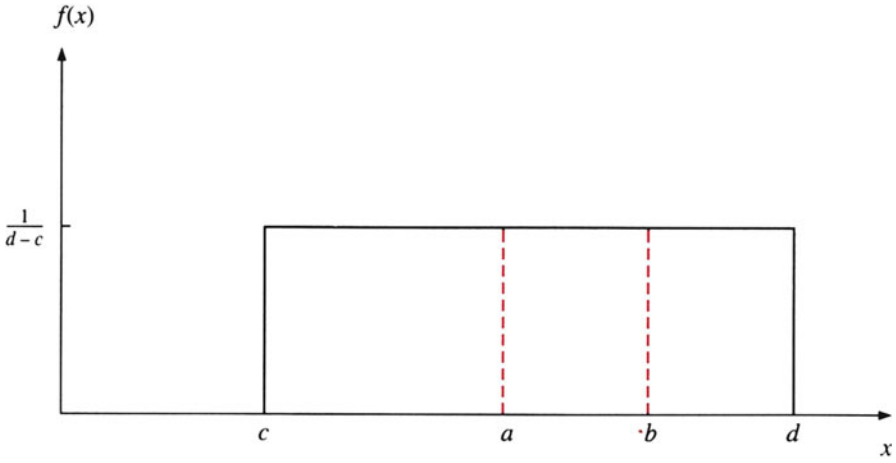


Fig. 7.3 The uniform probability distribution

so-called *cumulative uniform density function*. The density and cumulative density functions for this random variable can be defined as

$$f(X) = \begin{cases} \dfrac{1}{d-c} & \text{if } c \le X \le d \\ 0 & \text{elsewhere} \end{cases} \tag{7.3a}$$

$$F(X \le x) = \begin{cases} 0 & x<c \\ \dfrac{x-c}{d-c} & c \le x \le d \\ 1 & x>d \end{cases} \tag{7.3b}$$

where $X = x$ is a continuous random variable that represents a point in the interval $c \le x \le d$, as described in Fig. 7.3. Any one value of a uniform random variable is as likely to occur as any other, so the distribution is evenly spread over the entire region of possible values. In Chap. 9, we will discuss the *uniform distribution* further.

If there exist two points, a and b, between c and d in Fig. 7.3, the probability of x being between a and b for a uniform distribution can be calculated as follows: first, substituting $x = a$ and $x = b$ into Eq. 7.3b, we get

$$F(a) = \frac{a-c}{d-c}$$
$$F(b) = \frac{b-c}{d-c}$$

Then, substituting both $F(b)$ and $F(a)$ into Eq. 7.2, we obtain the probability of x being between a and b:

$$P(a < X < b) = \frac{b-a}{d-c}$$

For other types of continuous random variables, the calculation of Eq. 7.2 generally requires knowledge of calculus.[1] For example, if $c = 10$, $d = 20$, $a = 15$, and $b = 19$, then $P(a < X < b) = (19 - 15)/(20 - 10) = .30$.

7.3 The Normal and Standard Normal Distribution

7.3.1 The Normal Distribution

The *normal distribution* is the most widely used continuous density distribution in statistics. Many random variables have been found to be normally distributed, including measurements of weight, height, age, time, snowfall, yields, dimension, and other measures of interest to managers in both the public and private sectors.[2] When attempting to make an assertion about a population by using sample information, a major assumption we often make is that the population has a normal distribution.

From Fig. 7.4, it is obvious that the normal distribution is centered on its mean. Because the distribution is symmetric, the mean and median also occur at the same point. In addition, the bell-shaped normal curve has a single peak; it is unimodal. The *normal probability density function* (PDF) for a normal variable X gives the height of an observation such as cd in Fig. 7.4.[3]

---

[1] The probability that X lies between a and b for a continuous variable can be defined as

$$P(a < X < b) = \int_a^b f(x)\,dx = F(b) - F(a) \qquad \text{(A)}$$

where $f(x)$ represents the PDF of a continuous random variable X being valued at x.

From integral calculus, we know that the integration ($\int$) for the continuous case is the counterpart of the summation ($\sum$) in the discrete case. From Fig. 7.2 and Eq. A, we know that the probability for a continuous PDF is represented by the area bounded by the curve whose value at x is $f(x)$, by the x-axis, and by the lines $x = a$ and $x = b$. Discussion of areas under the continuous PDF and of the mean and variance of a continuous variable can be found in Appendix 1.

[2] Karl F. Gauss (1777–1855) discovered that the measurement of errors often follows a normal distribution.

[3] The PDF of a normal random variable can be defined as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

where $\pi = 3.14159$, $e = 2.71828$, and $\mu(-\infty < \mu < \infty)$ and $\sigma^2(0 < \sigma^2 < \infty)$ are the mean and variance of the normal random variable X. To graph the normal curve, we must know the numerical values of μ and σ^2.

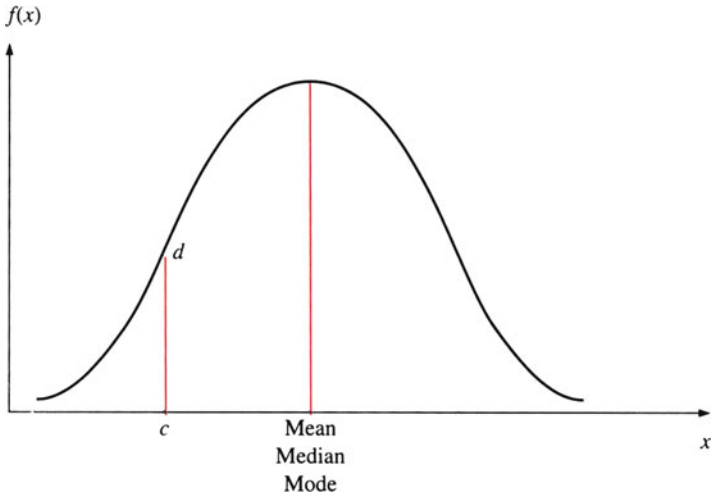


Fig. 7.4 Normal distribution

Normally distributed populations with different shapes may nevertheless have many characteristics in common. The factor that determines shape is standard deviation: the larger the standard deviation, the wider the curve. Figure 7.5 shows MINITAB results for three normal distributions with mean 0 and three different standard deviations. Note that the mean μ and standard deviation σ completely characterize the normal PDF.

7.3.2 Areas Under the Normal Curve

7.3.2.1 Measuring the Area Under a Normal Curve

No matter what the values of μ and σ for a normal probability distribution, the total area under the normal curve is 1.00, so we may think of areas under the curve as representing probability. Table 7.2 shows how the area under the curve within a certain interval can be determined mathematically.[4]

There is no closed-form expression for $P(a < X < b) = \int_a^b f(x)\,dx$ for the normal probability distribution. However, the value of the definite integral can be obtained by numerical approximation procedures. The areas in Table A3 in Appendix A were obtained by using such a procedure (see Appendix 1).

[4] It is virtually impossible to capture all observations under the curve because, theoretically, the tails continue indefinitely in both directions, never touching the horizontal axis. Integrating the probability density function over the range from $-\infty$ to $+\infty$ would yield the total area of a normal distribution, which is equal to 1.

```
MTB > SET C1
DATA> -3:3/0.1
DATA> PDF C1 C2;
SUBC> NORMAL 0 1.
MTB > PDF C1 C3;
SUBC> NORMAL 0 5.
MTB > PDF C1 C4;
SUBC> NORMAL 0 2.
MTB > GPRO
* NOTE  * Professional Graphics are enabled.
          Standard Graphics are disabled.
          Use the GSTD command to enable
          Standard Graphics.
MTB > Plot C2*C1 C3*C1 C4*C1;
SUBC>   Connect;
SUBC>     Type 1;
SUBC>     Color 1;
SUBC>     Size 1;
SUBC>   Overlay;
SUBC>   Axis 1;
SUBC>     Label "X";
SUBC>   Axis 2;
SUBC>     Label "PROBABILITY".
```

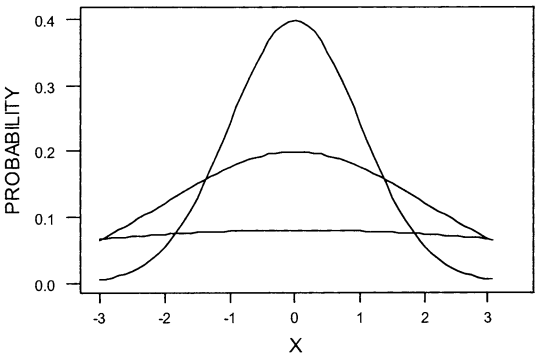


Fig. 7.5 MINITAB results of normal distributions with different standard deviations

Table 7.2 Using the mean and standard deviation to determine the area of a normal distribution

Approximately 68.26 % of the area (probability) under the normal curve lies between $-\sigma$ and $\mu + \sigma$.
Approximately 95.45 % of the area (probability) under the normal curve lies between $\mu - 2\sigma$ and $\mu + 2\sigma$.
Approximately 99.73 % of the area (probability) under the normal curve lies between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Table 7.3 Mean and standard deviation of EPS

With leverage	Without leverage
$\mu_{EPS} = \$1.98$	$\mu_{EPS} = \$1.80$
$\sigma_{EPS} = \$0.32$	$\sigma_{EPS} = \$0.20$

Example 7.4 The Normal Distribution: An Application to EPS. Firms frequently use debt to fund various projects. The overall level at which a company employs debt throughout its operations directly affects the expected level and range of its earnings per share (EPS). This is because of the increased risk to which debt exposes the firm within the capital markets.[5] A firm that employs debt is said to be leveraged. The more debt it uses, the higher the firm is leveraged. A firm that does not use any debt financing is said to be without leverage.

If we calculate the means and standard deviations for a hypothetical firm with and without leverage, we can estimate the interval of the possible EPS in the future. The means and standard deviations of the firm are listed in Table 7.3. Figure 7.6 shows the distributions in terms of the parameter values given in Table 7.3.

When we say that EPS is normally distributed, we mean that as the number of observations becomes very large, graphing them yields a normal curve. We can predict EPS intervals for the firm from the information given in Table 7.3 by using

[5] Increased risk results because interest and principal payments on debt represent legal obligations. Because stock represents ownership in a company, dividends are *not* a legal obligation of the firm.

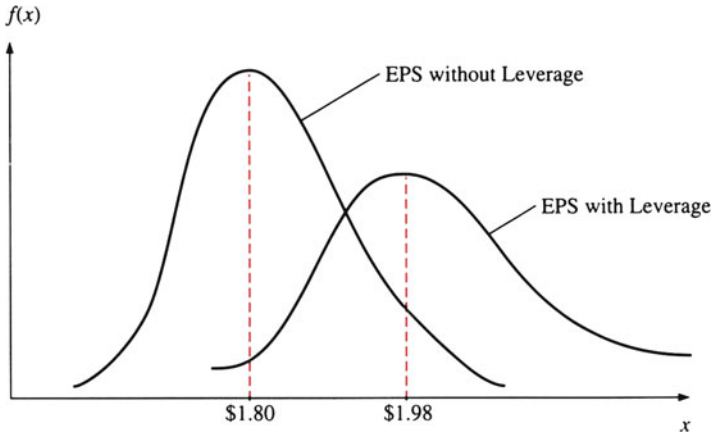


Fig. 7.6 EPS distributions

Table 7.4 Interval statements about the EPS for a hypothetical firm

Chance	With leverage	Without leverage
68 %	$\$1.66 \leq X \leq \$2.30$	$\$1.60 \leq X \leq \$2.00$
95.5 %	$\$1.34 \leq X \leq \$2.62$	$\$1.40 \leq X \leq \$2.20$
99.7 %	$\$1.02 \leq X \leq \$2.94$	$\$1.20 \leq X \leq \$2.40$

the empirical rule described in Table 7.2. Over a large number of observations, x percent will lie within a specified interval that we can determine via the mean and standard deviation. Table 7.4 shows the results of predictions about the firm's EPS. In Table 7.4, X is the EPS for the firm. For example, the last interval statement implies that without leverage, approximately 99.7 % of EPS should lie between \$1.20 and \$2.40.

7.3.2.2 The Standard Normal Distribution and the *Z* Statistic

There is an infinitely large number of normal curves—one for each pair of values for μ and σ. It is neither possible nor necessary to have different tables for every possible normal curve. The *standard normal distribution* is a transformation of the normal distribution. In the standard normal curve, $\mu = 0$ and $\sigma = 1$. This standard normal curve can be displayed in terms of Z scores (presented in Sect. 4.5) as indicated in Fig. 7.7.

The standard normal curve helps simplify the calculation of probabilities for normally distributed populations. Because not all normally distributed random variables have $\mu = 0$ and $\sigma = 1$, we need to transform the variable so that $\mu = 0$ and $\sigma = 1$. We do this by using the Z score, which is calculated as follows:

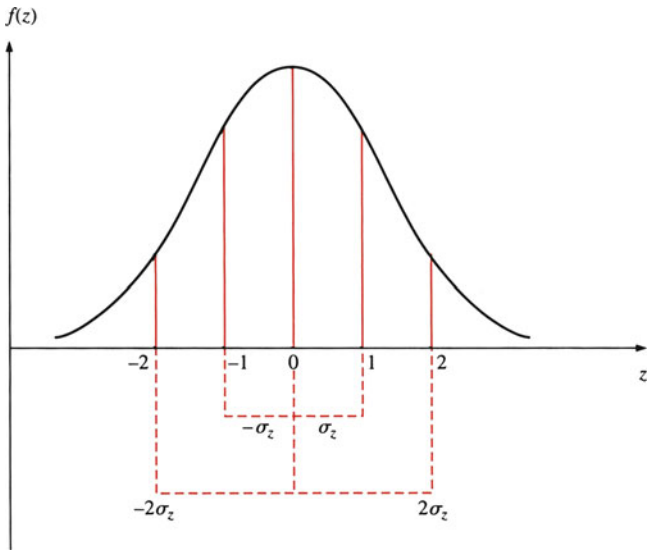


Fig. 7.7 Normal probability distribution with $\mu = 0$ and $\sigma = 1$

$$Z = \frac{X - \mu}{\sigma} \tag{7.4}$$

The Z score represents the distance, or deviation, between a given value of the continuous random variable X and its mean μ in standard units. With this information in hand, we can construct the standard normal area table as presented in Table A3 in Appendix A to calculate the area under the curve associated with the value of Z. It is important to note that for any positive value of Z, we are looking at only half the curve. We must therefore add .5 to that value to find the total area under the curve at or below that point.

7.3.3 How to Use the Normal Area Table

Assume that the IQs of undergraduate students at your school are normally distributed with $\mu = 120$ and $\sigma = 15$. What proportion of these undergraduates have an IQ between 120 and 142.5? In this case, we have to find the area of the shaded portion in Fig. 7.8.

To use the normal area table, we need to calculate the 2 value for $X = 142.5$:

$$Z = \frac{X - \mu}{\sigma} = \frac{142.5 - 120}{15} = 1.5$$

This implies that the value 142.5 lies 1.5 standard deviations above the mean. Using this information and the normal area table (Table A3 of Appendix A), we find that the corresponding portion in the table is .4332, as indicated in Fig. 7.8.

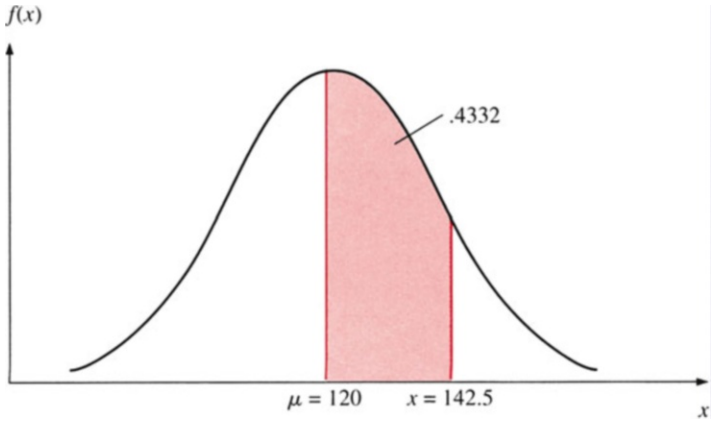


Fig. 7.8 Normal distribution for student IQs with interval between 120 and 142.5 in *shaded area*

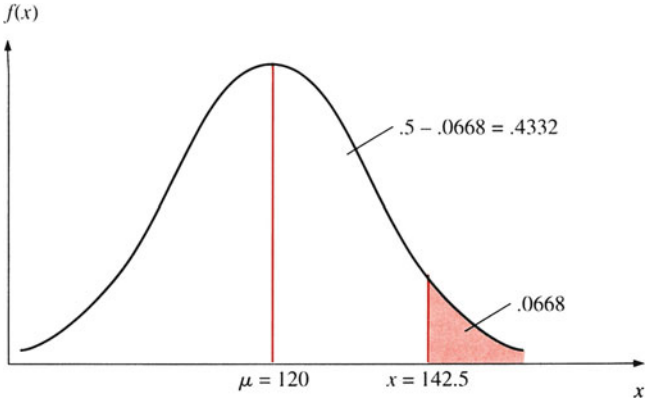


Fig. 7.9 Normal distribution for student IQs with intervals above 142.5 in *shaded area*

Note that the portions in the normal area table show the area under the upper tail of the normal curve. Because the total area under the normal curve is 1.0, each half is .5. Hence, the area that we seek in Fig. 7.9 is .5 = .0668 = .4332. The sought area is 43.32 % of the total area. From the probability concepts discussed in Chap. 5, we can write the event "IQ between 120 and 142.5" as *E,* and we conclude that $P(E) =$ .4332. In other words, 43.32 % of the undergraduate students at your school have an IQ between 120 and 142.5.

Alternatively, the MINITAB program can be used to calculate the percentage of undergraduate students at your school who have IQ scores between 120 and 142.5, as indicated here:

```
MTB > SET INTO C1
DATA> 142.5 120
DATA> END
MTB> CDF C1;
```

```
SUBC> NORMAL 120 15.
142.5000 0.9332
120.0000 0.5000
MTB> PAPER
```

From the example, we know that 93.32 % of the students at your school have IQ scores of 142.5 or below and 50 % of the students have IQ scores of 120 or below. By subtracting 50 % from 93.32 %, we obtain 43.32 %.

The marketing manager of a chain of supermarkets needed to know the weekly sales of extra large eggs. He asked one of his staff to do a survey over a 25-week period. The survey revealed that the weekly sales of extra large eggs were normally distributed, with a mean of 743 cartons and a standard deviation of 254 cartons (*Journal of Marketing Research,* August 1984).

From this information, we can calculate the probability that a supermarket will sell between 550 and 850 cartons of extra large eggs in a randomly selected week as:

$$\begin{aligned} P\{550 < X < 850\} &= P\left\{\frac{550-743}{254} \le X \le \frac{850-743}{254}\right\} \\ &= P\{z \le .42\} + P\{z \le -.76\} \\ &= P\{z \le .42\} + P\{z \ge .76\} \\ &= .1628 + .2764 = .4392 \qquad \text{(From Table A3)} \end{aligned}$$

Example 7.5 Determining Daily Donut Demand (in Dozens). A Dunkin' Donuts shop located in New Brunswick, New Jersey, sells dozens of fresh donuts. Any donuts remaining unsold at the end of the day are either discarded or sold elsewhere at a loss. The demand for the Dunkin' Donuts at this shop has followed a normal distribution with $\mu = 50$ dozen and $\sigma = 5$ dozen. How many dozen donuts should this Dunkin' Donuts shop make each day so that it can meet the demand 95 % of the time?

Let the random normal variable X represent the demand for fresh Dunkin' Donuts (measured in dozens). To meet the demand 95 % of the time, the Dunkin' Donuts shop must determine an amount—say, A dozen—such that

$$P(X \le A) = .95$$

Similarly to the student IQ case, we can express this probability statement in terms of Z:

$$P\left(\frac{X-\mu}{\sigma} \le \frac{A-\mu}{\sigma}\right) = P\left(Z \le \frac{A-\mu}{\sigma}\right) = .95$$

Because we know that $\mu = 50$ and $\mu = 5$, we can rewrite the probability statement as

$$P\left(Z \le \frac{A-50}{5}\right) = .95$$

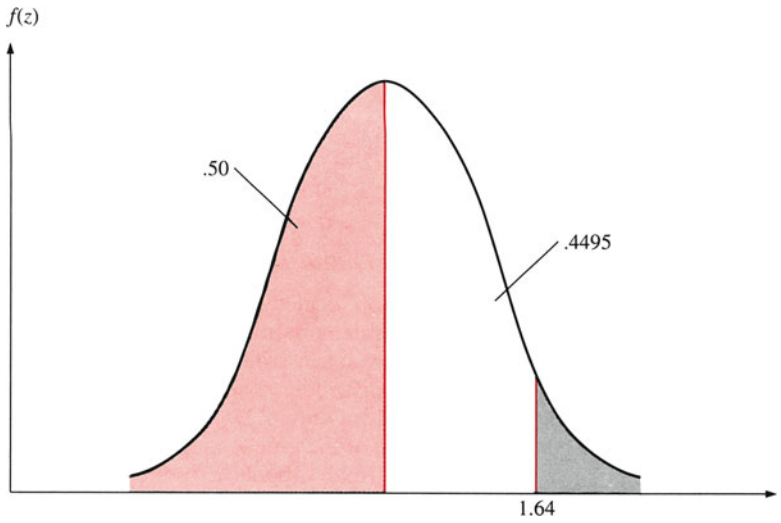


Fig. 7.10 $P(Z \leq 1.64)$

In accordance with table A3 in Appendix A, a Z curve having an area to the left equals .95, as shown in Fig. 7.10. From Fig. 7.10, we know that

$$P(0 \leq Z \leq 1.64) = .4495 \cong .45$$

which means that

$$P(Z \leq 1.64) = .5 + .45 = .95$$

Thus

$$\frac{A - 50}{5} = 1.64$$
$$A = 58.2$$

and

$$A = 50 + 8.20 = 58.20 \text{ dozen}$$

To be conservative, we round this value up to 59 dozen and assume that any occasional surplus will be welcome at a nearby shelter for the homeless. By stocking 59 dozen donuts each day, the Dunkin' Donuts shop will meet the demand for donuts 95 % of the time.

7.4 The Lognormal Distribution and Its Relationship to the Normal Distribution (Optional)

7.4.1 The Lognormal Distribution

Before we discuss the *lognormal distribution,* we must briefly review and expand on three topics covered in Chap. 4: mean, variance, and skewness.[6] For a continuous random variable, we can generally calculate the mean, variance, and skewness. Values of these parameters affect the shape of a distribution. Lognormally distributed random variables are related to the normally distributed continuous variables, but normally distributed random variables have zero skewness, whereas lognormally distributed continuous random variables have positive skewness.

If a continuous random variable Y is normally distributed, then the continuous variable X defined in Eq. 7.5 is lognormally distributed:

$$X = e^{Y} \tag{7.5}$$

By performing a logarithmic transformation on this variable X, we obtain a normally distributed variable Y:

$$\ln(X) = \ln(e^{Y}) = Y$$

where In denotes the natural logarithm and e is a constant approximately equal to 2.71828. Lognormal continuous random variables have the following properties:

$$\text{Mean}: \; E[\ln(X)] = E(Y) = \mu$$
$$\text{Variance}: \; \text{Var}[\ln(X)] = \text{Var}(Y) = \sigma^2$$

Our discussion in the next section of the mean and variance for a lognormal variable X is based on these relationships.

7.4.2 Mean and Variance of Lognormal Distribution

Because of the relationship between X and Y indicated in Eq. 7.5, the mean and variance of variable X can be denned as follows[7]:

[6] This section can be omitted without affecting the continuity of the text. Further discussion on the lognormal distribution can be found in Aitchison, J., Brown, J.A.C.: The Lognormal Distribution with Special Reference to Its Uses in Economics. Cambridge University Press, London (1957).

[7] The density function of a lognormal distribution can be defined as

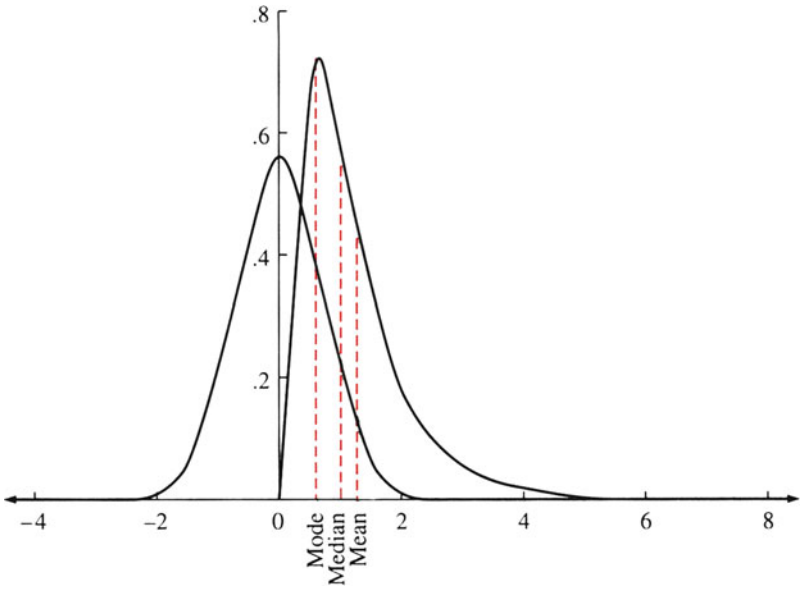


Fig. 7.11 Frequency curves of the normal and lognormal distributions (*Source:* Nelson, C.R.: Applied Time Series Analysis, p. 164. Holden-Day, Oakland (1973))

$$\mu_X = e^{\mu + 1/2\sigma^2} \tag{7.6}$$

$$\sigma_X^2 = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right) \tag{7.7}$$

where $\mu = E[\ln X]$, $\sigma^2 = \text{Var}[\ln X]$, and $e = 2.71828$.

Equations 7.6 and 7.7 indicate that both mean and variance of a lognormal variable are functions of the mean and variance of a normal variable. The normal and the corresponding lognormal frequency curves are illustrated in Fig. 7.11. Note that the mean of the lognormal is larger than the mode of the lognormal, because it is a positively skewed distribution. In addition, the shape of a lognormal is affected by[7] the values of both mean μ and variance σ^2, as indicated in Figs. 7.12 and 7.13. Furthermore, the lognormal distribution differs from the normal distribution in that its mean, median, and mode are not identical.

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\ln x - \mu)^2\right], \quad x > 0$$

This is similar to the density function of a normal distribution as defined in footnote 3 of this chapter. Mean, variance, and skewness of the lognormal distribution will be discussed and derived in Sect. 9.7.

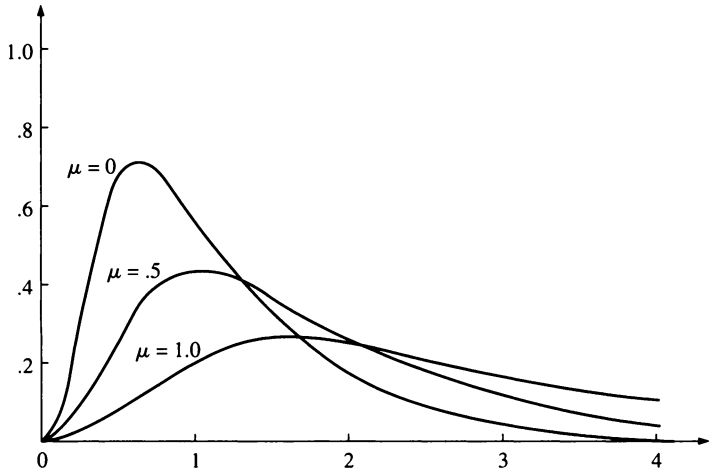


Fig. 7.12 Frequency curves of the lognormal distribution for three values of μ from the parent normal (*Source:* Nelson, C.R.: Applied Time Series Analysis, p. 164. Holden-Day, Oakland (1973))

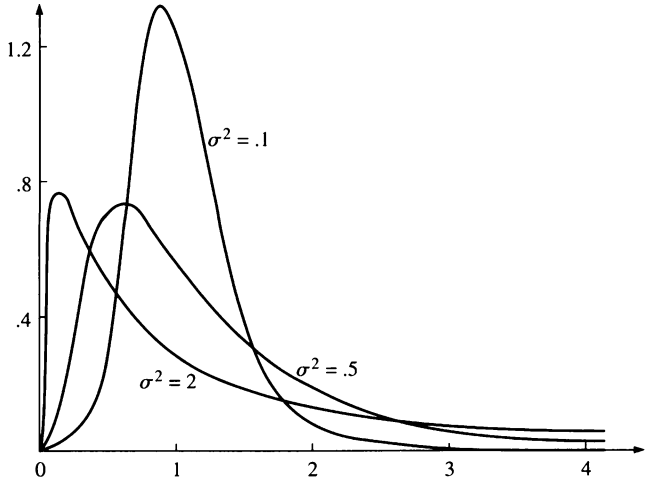


Fig. 7.13 Frequency curves of the lognormal distribution for three values of σ^2 (*Source:* Nelson, C.R.: Applied Time Series Analysis, p. 165. Holden-Day, Oakland (1973))

In surveys of household income and in the examination of consumer behavior, the lognormal distribution is useful in the income distribution analysis.[8] In addition, the lognormal distribution is more suitable than the normal distribution for cost–volume–profit analysis, which is discussed in Application 7.2. Furthermore,

[8] This is because household income is generally lognormally distributed. See Aitchison and Brown (1957), Chapters 11 and 12.

in the option pricing model discussed in Appendix 2, it is assumed that stock price per share is lognormally distributed (see Eq. 7.35).

To show how Eqs. 7.6 and 7.7 can be used to calculate the mean and standard deviation of a lognormal distribution, we let X represent the stock price per share of JNJ as presented in Table 2.3 in Chap. 2. Then, we can calculate $E(\ln(X)) = E(Y) = 4.0749$ $\text{Var}[\ln(X)] = \text{Var}(Y) = \sigma^2 = .08995$ by MINITAB as shown here. Substituting $E(Y) = \mu = 4.0749$ and $\text{Var}(Y) = \sigma^2 = .08995$ into Eqs. 7.6 and 7.7, we obtain

$$\mu_X = e^{4.11988} = 61.5516$$
$$\sigma_X^2 = (61.5516)^2(e^{.08995} - 1) = 356.5814.$$

```
MTB >                                   SET INTO C1
DATA>                                   57.63 78.50 62.88 53.75
                                         50.00 45.00 38.50 62.38
                                         74.38 78.38
DATA>                                   61.38 83.50 42.25 34.38
                                         28-88 32.25 54.88 42.13
                                         52.88 58.00
DATA>
                                         END
MTB >
                                         MEAN C1
Column Mean
  Mean of C1 = 54.597
MTB >
                                         STDEV C1
Column Standard Deviation
Standard deviation of C1 = 15.916
MTB >
                                         LET C2=LOGE(C1)
MTB >
                                         PRINT C2
Data Display
C2
4.05404 4.36310                 4.14123
        3.98434                  3.91202
        3.80666                  3.65066
        4.13324
4.30919 4.36157                 4.11708
        4.42485                  3.74360
        3.53747                  3.36315
        3.47352
4.00515 3.74076                 3.96803
        4.06044
```

```
MTB >                                   MEAN C2
Column Mean
  Mean of C2 = 3.9575
MTB > STDEV C2
Column Standard Deviation
Standard deviation of C2 = 0.30423
```

7.5 The Normal Distribution as an Approximation to the Binomial and Poisson Distributions

In Chap. 6, we discussed binomial and Poisson distributions. Recall that when we were interested in deriving the cumulative probability from the binomial distribution, the computations could be quite burdensome. For example, if we toss a coin 100 times to test the probability that the number of heads will be 50 or less, we need to compute $P(X = 0)$, $P(X = 1)$, $P(X = 2)$, . . . $P(X = 50)$. An analogous situation arises for the Poisson distribution. Fortunately, it is possible to reduce the job of computation greatly by approximating the binomial or Poisson distribution with a normal distribution.

7.5.1 Normal Approximation to the Binomial Distribution

As the sample size gets large, we can use a normal distribution to approximate the binomial distribution. For an experiment that does n independent trials each having probability of success p, the distribution of the number of successes, X, is binomial and has the following mean and variance:

$$\text{Mean}: \quad E(X) = \mu = np \tag{7.8}$$

$$\text{Variance}: \quad \text{Var}(X) = np(1-p) \tag{7.9}$$

From Eq. 7.4, we substitute for the mean μ and standard deviation σ and get the following expression for the Z statistic:

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \tag{7.10}$$

We say that the distribution of the random variable Z is approximately standard normal. This approximation works well when $np > 5$ and $n(1-p) > 5$. Because X stands for the number of successes of the binomial trials, we can now determine the probability that the actual number of successes will lie within a certain interval.

If the range we wish to examine is between a and b, inclusive, then we may obtain the following probability:

$$P(a \leq X \leq b) = P\left(\frac{a - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b - np}{\sqrt{np(1-p)}}\right) \tag{7.11}$$

Example 7.6 Using a Normal Distribution to Approximate a Binomial Distribution. Suppose a very bumpy conveyor belt in a brewery transports beer bottles from the point where they are capped to the point where they are boxed for shipping. Furthermore, let us suppose there is a 16 % chance that each beer bottle will fall off the conveyor belt. In 1 h, exactly 1,000 beer bottles travel from one end of the belt to the other. Because n is large, we can make probability statements about whether the actual number of bottles, X, that fall off the conveyor will be within a certain range. Using Eqs. 7.8 and 7.9, we get the following results for a binomially distributed random variable with the mean and variance shown:

$$E(X) = np = 160$$

$$Var(X) = np(1-p) = 134.4$$

Suppose we wish to know the probability that the actual number of beer bottles that fall off the conveyor belt will be between 142 and 185. Using Eq. 7.11 yields

$$\begin{aligned} P(142 \leq X \leq 185) &= P\left(\frac{142 - 160}{\sqrt{134.4}} \leq Z \leq \frac{185 - 160}{\sqrt{134.4}}\right) \\ &= P(-1.553 \leq Z \leq 2.156) \end{aligned}$$

Now we can use the values of the Z statistic and the standard normal distribution table to compute the probability. We calculate the area beneath the curve between these two numbers. Then we let the symbol F_z represent the value of cumulative probability as taken from the standard normal distribution table. Then

$$\begin{aligned} P(-1.55 \leq Z \leq 2.16) &= F_Z(2.16) - F_Z(-1.55) \\ &= F_Z(2.16) - [1 - F_Z(1.55)] \\ &\quad \text{Because } F_Z(-w) = 1 - F_Z(w) \\ &= .9846 - [1 - .9394] \\ &= .9240 \end{aligned}$$

Thus, there is a 92.4 % chance that the number of beer bottles that fall off the conveyor belt during the period will be between 142 and 185.

This example illustrates how the normal distribution can be used to approximate the binomial distribution. Further applications of the normal distribution are given in Appendix 2.

7.5.2 Normal Approximation to the Poisson Distribution

Recall from Sect. 6.7 that the Poisson random variable measures the probability of X occurrences of some event in the time interval between 0 and t. Therefore, this distribution measures successes when they occur within specified units of time. Recall the Poisson probability function:

$$P(X = x) = e^{-\lambda}\lambda^{x}/x! \qquad \text{for } x = 0,\ 1,\ 2,\ 3,\ \ldots \text{ and } \lambda > 0 \tag{6.16}$$

where λ is the average number of successes in the unit of time and e is the base of the natural logarithms (2.71828). The mean and variance of this distribution are

$$\text{Mean}: \qquad E(X) = \mu = \lambda \tag{7.12}$$

$$\text{Variance}: \qquad \text{Var}(X) = \sigma^2 = \lambda \tag{7.13}$$

Note that both the variance and the mean are equal to λ.

The Poisson probabilities can be approximated by the normal distribution when the sample size is large—say, greater than 30. To calculate Poisson probabilities in this manner, we can develop the Z statistic by substituting for the mean and variance in Eq. 7.4 as follows:

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \tag{7.14}$$

Now we can examine the probability that the number of successes is within a certain range.

$$P(a \le X \le b) = P\left(\frac{a - \lambda}{\sqrt{\lambda}} \le Z \le \frac{b - \lambda}{\sqrt{\lambda}}\right) \tag{7.15}$$

Example 7.7 Using a Normal Distribution to Approximate a Poisson Distribution. In Example 6.16, we examined the average number of customers entering a bank in a 10-min period. Now let's assume that in a 20-min period, we have an average of 50 customers instead of 5; X is still a Poisson random variable. This change is made so that normal approximation will hold. Eqs. 7.12 and 7.13 reveal that the mean and variance of this Poisson random variable are both 50.

Suppose we wish to find the probability that the number of people entering the bank in a 20-min period will be between 42 and 57, inclusive. We can calculate this probability by using the standard normal distribution and its related Z statistic:

$$P(42 \le X \le 57) = P\left(\frac{42-50}{\sqrt{50}} \le Z \le \frac{57-50}{\sqrt{50}}\right) = P(-1.13 \le Z \le .99)$$

Now the F_z's are computed in the same way as in Example 7.6. Then

$$\begin{aligned} P(-1.13 \le Z \le .99) &= F_Z(.99) - F_Z(-1.13) \\ &= .8389 - (.1292) \\ &= .7097 \end{aligned}$$

There is a 70.97 % chance that the number of customers who arrive in a 20-min period will fall between 42 and 57, inclusive.

7.6 Business Applications

Application 7.1 Analyzing Earnings per Share and Rates of Return. In financial analysis, historical data are often used to forecast future values. Table 7.5 presents data on earnings per share and rates of return on stock in Johnson & Johnson and Merck over a 20-year period (1990–2009). Mean, standard deviation, and skewness estimates for EPS and rates of return are presented in Table 7.6.

By analyzing past data, and by assuming that the data on JNJ and MRK are distributed normally, we can make interval statements about EPS and return for JNJ and MRK, as indicated in Table 7.7.

First consider the return on JNJ and MRK stock. The average return on JNJ (see Table 7.6) is higher than that on MRK. However, the standard deviation measure presented in Table 7.6 makes it clear that Merck's stock is less volatile than Johnson and Johnson's. An investor seeking a higher return might choose Johnson and Johnson but, in doing so, would incur a greater risk of losing money.

Now we analyze the mean and standard deviation of EPS presented in Table 7.6. EPS is an absolute measure, so the EPS data for MRK and JNJ are not directly comparable; one share of Merck stock does not cost the same as one share of Johnson & Johnson stock. Therefore, EPS are earnings on different amounts of investment per share. The problem can be resolved by comparing the *coefficient of variation* (CV) of EPS for Merck and Johnson & Johnson. Recall that the coefficient of variation, which we discussed in Chap. 4, divides the standard deviation by the mean and gives an indication of relative volatility. Substituting the related data of Table 7.6 into Eq. 4.11, we obtain the coefficients of variation of EPS for both Merck and Johnson & Johnson:

$$\text{CV(JNJ)} = .8694/3.0375 = 0.2862 \qquad \text{CV(MRK)} = 1.1640/3.0898 = 0.3767$$

Table 7.5 EPS and rates of return for JNJ and MRK

	EPS		Rate of return	
Year	*JNJ*	*MRK*	*JNJ*	*MRK*
1990	3.38	4.51	1.778	1.332
1991	4.30	5.39	1.835	1.532
1992	1.54	1.70	1.582	1.216
1993	2.71	1.86	1.624	.973
1994	3.08	2.35	1.566	1.270
1995	3.65	2.63	1.809	1.515
1996	2.12	3.12	1.807	1.600
1997	2.41	3.74	1.999	1.475
1998	2.23	4.30	1.364	1.685
1999	2.94	2.45	1.771	1.285
2000	3.39	2.90	2.164	1.375
2001	1.83	3.14	2.296	1.123
2002	2.16	3.14	1.683	1.199
2003	2.39	3.03	1.710	1.205
2004	2.83	2.61	1.962	1.147
2005	3.46	2.10	2.485	1.582
2006	3.73	2.03	1.199	1.197
2007	3.63	1.49	1.510	1.227
2008	4.57	3.64	1.649	1.348
2009	4.40	5.68	1.820	1.805

Table 7.6 Mean, standard deviation, and skewness estimates for EPS and rates of return

	EPS		Rate of return	
Year	*JNJ*	*MRK*	*JNJ*	*MRK*
Mean	3.0375	3.0898	1.7806	1.3545
Std. Dev.	0.8694	1.1640	0.3010	0.2124
Skewness	0.1257	0.8487	0.5253	0.4382

Table 7.7 Probability distribution for EPS and return

	EPS	
Chance	*JNJ*	*MRK*
68.3 %	$2.1681 \leq X \leq 3.9069$	$1.9258 \leq X \leq 4.2538$
95.4 %	$1.2987 \leq X \leq 4.7763$	$0.7618 \leq X \leq 5.4178$
99.7 %	$0.4293 \leq X \leq 5.6457$	$-0.4022 \leq X \leq 6.5818$
	Rate of return	
68.3 %	$1.4796 \leq X \leq 1.5669$	$1.1421 \leq X \leq 1.5669$
95.4 %	$1.1785 \leq X \leq 1.7793$	$0.9297 \leq X \leq 1.7793$
99.7 %	$0.8776 \leq X \leq 1.9917$	$0.7173 \leq X \leq 1.9917$

Comparing these two coefficients of variation, we see that Merck's EPS is much more volatile than JNJ's. Therefore, the risk-averse investor might prefer JNJ stock to Merck.

Application 7.2 Cost–Volume–Profit Analysis Under Uncertainty: The Normal Versus the Lognormal Approach. Cost–volume–profit (CVP) analysis is one of

the most important concepts in accounting, economics, finance, marketing, and production management. The total profit w of a firm can be defined as

$$w = TR - TC = Q(P - V) - F \tag{7.16}$$

where

TR = total revenue
TC = total cost
Q = unit sales
P = price per unit
V = variable cost per unit
$(P - V)$ = contribution margin per unit
F = fixed cost

This model can be analyzed in terms of the certainty approach or the uncertainty approach. Under certainty analysis, we assume that future Q, P, and V are known for sure and that, accordingly, the specified future total profit will occur with 100 % certainty.

Uncertainty analysis is more complicated and (not surprisingly) less certain. Hilliard and Leitch (1975) have suggested two alternative assumptions[9]:

1. Q is not known for certain, and it is normally distributed with mean μ_q and variance. σ_q^2. Thus, the random variable w is normally distributed with mean μ_w and variance σ_w^2 as follows:

$$\begin{aligned} \mu_w &= \mu_q(P - V) - F \\ \sigma_w^2 &= \sigma_q^2(P - V)^2 \end{aligned} \tag{7.17}$$

Following Hilliard and Leitch (1975), we suppose that μ_q = 5,000 units, σ_q = 400 units, price = \$3,000/unit, variable cost = \$l,750/unit, and fixed costs = \$5,800,000. Thus, μ_w = \$450,000 and σ_w= \$500,000. We can calculate the probability of a profit greater than \$200,000 (A) by using Table A3 in Appendix A:

$$\begin{aligned} P(w > \$200,000) &= 1 - P(w \le \$200,000] \\ &= 1 - F_w[(A - \mu_w)/\sigma_w] \\ &= 1 - F_w\left(\frac{200,000 - 450,000}{500,000}\right) \\ &= 1 - F_w(-.5) = 1 - .3085 = .6915 \end{aligned}$$

[9] Hilliard, J.E., Leitch, R.A.: Cost-volume-profit analysis under certainty: A lognormal approach. Acc. Rev., 69–80 (1975 January).

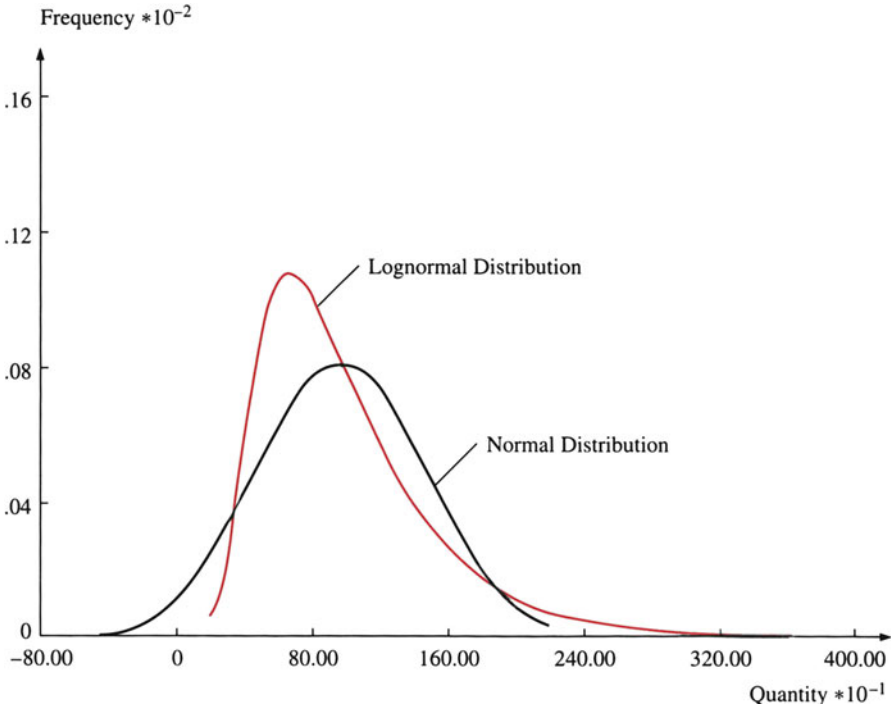


Fig. 7.14 Comparison of normal and lognormal distributions: Coefficient of variation = .50 (*Source:* Hilliard, Leitch, CVP under uncertainty. Acc. Rev. p. 71, January (1975))

This implies there is a 69.2 % chance that the firm's profit will exceed $200,000. Similarly, we can calculate the probability of meeting the break-even point by setting w equal to 0.

2. Q is not known for certain, and it is lognormally distributed. Following Eqs. 7.6 and 7.7, we can define the mean and variance of Q as

$$\mu_q = E(Q) = e^{\mu + 1/2\sigma^2}$$
$$\sigma_q^2 = \mathrm{Var}(Q) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

where $\mu = E(\ln Q)$ and $\sigma^2 = \mathrm{Var}(\ln Q)$.

Logical assumptions for a CVP model would require that sales be nonnegative. In addition, it would be somewhat surprising if the distribution of sales (Q) were perfectly symmetric about its mean. Thus, assuming the lognormal distribution might be more suitable than assuming the normal.

The relationships between the normal and lognormal distributions under conditions of large and small coefficients of variation are presented in Figs. 7.14

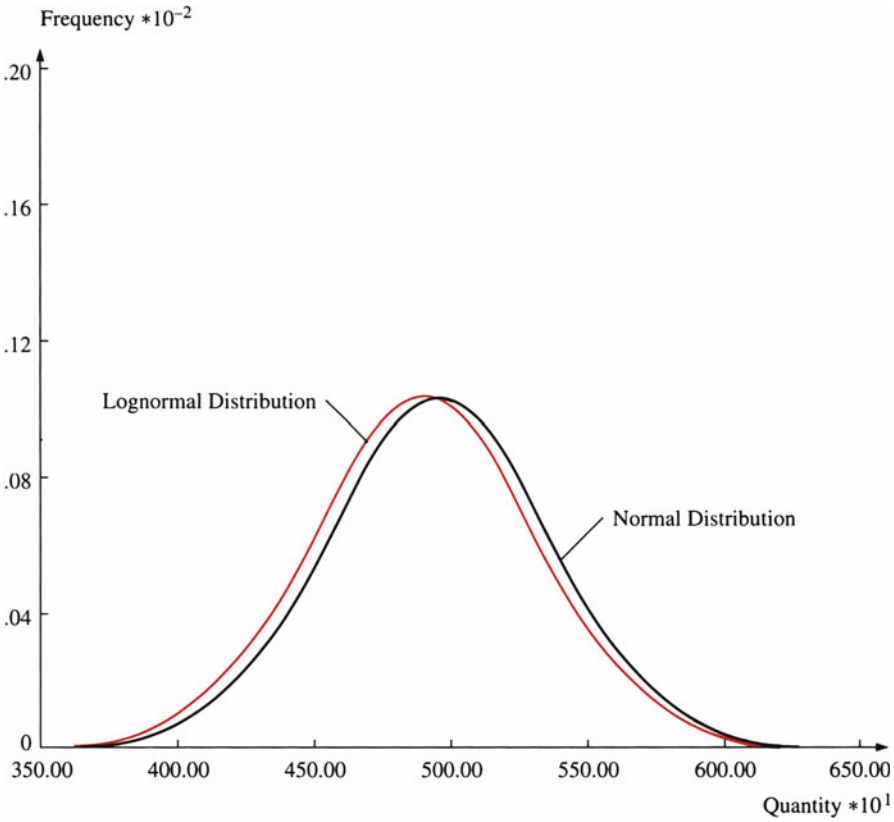


Fig. 7.15 Comparison of normal and lognormal distributions: Coefficient of variation = .08 (*Source:* Hilliard, Leitch, CVP under uncertainty. Acc. Rev. p. 72, January (1975))

and 7.15, respectively. In Fig. 7.14, both the normal and the lognormal distribution have identical means, identical variances, and a coefficient of variation of .50. This relatively large coefficient of variation emphasizes the difference between the normal and lognormal distributions. The figure indicates that the major differences between these two distributions are the skewness of the lognormal and the possibility of the occurrence of negative values for the normal distribution. In Fig. 7.15, the distributions are nearly coincident, as we expect for small coefficients of variation (CV = .08). The important observation, however, is that the lognormal assumption is an intuitive choice for the CVP model inputs, regardless of the values of the coefficient of variation.

Under the lognormal distribution assumption, Hilliard and Leitch (1975) have derived the probability of a given level, say A, as

$$\begin{aligned} P[w > A] &= 1 - P[w \leq A] \\ &= 1 - F_w\{[\ln(A + F) - \mu_w]/\sigma_w\} \end{aligned} \tag{7.18}$$

where

$$\mu_w = \ln\left[\mu_q^2/\left(\sigma_q^2 + \mu_q^2\right)^{1/2}\right] + \ln(P - V)$$
$$\sigma_w^2 = \ln\left[(\sigma_q/\mu_q)^2 + 1\right]$$

Substituting the information used in case 1 into these two formulas, we obtain

$$\begin{aligned} \mu_w &= 2\ \ln(5,000) - \frac{1}{2}\ln[(400)^2 + (5,000)^2] + \ln(3,000 - 1,750) \\ &= 17.034 - 8.521 + 7.131 = 15.644 \\ \sigma_w^2 &= \ln\left[\left(\frac{400}{5,000}\right)^2 + 1\right] = .0064 \\ \sigma_w &= .08 \end{aligned}$$

Substituting A, F, μ_w, and σ_w into Eq. 7.18 yields

$$\begin{aligned} P[w{>}200,000] &= 1 - F_w\{[\ln(200,000 + 5,800,00) - 15.644]/.08\} \\ &= 1 - F_w[-.4625] \\ &= 1 - .3219 \\ &= .6781 \end{aligned}$$

The probability we get when we make the lognormal assumption is 67.81 %, which is lower than that in terms of the normal assumption, 69.15 %.

Application 7.3 Investment Decision Making Under Uncertainty. Professor Hillier (1963) suggested several easy and effective ways for a business firm to evaluate risky investment projects.[10] In one of his approaches, Hillier assumes that the *net cash inflow* from an investment to the firm in the tth future year after the investment is made is normally distributed.[11] He has also shown that the net present value (NPV) of a proposed investment is normally distributed with mean μ_{NPV} and

[10] Hillier, R.S.: The derivation of probabilities information for the evaluation of risky investments. Manage. Sci., 443–457 (April 1963). Section 21.8 and Appendix 4 (Chap. 21) will discuss this issue in further detail.

[11] Via Eq. 7.16, net cash inflow can be defined as net profit + depreciation.

variance $\sigma^2_{\text{NPV.}}$[12] Using the assumption that NPV is normally distributed, Hillier has provided an example of how management can evaluate the risk of an investment.

> Suppose that, on the basis of the forecasts regarding prospective cash flows from a proposed investment of \$10,000, it is determined that $\mu_p = \$1{,}000$ and $\sigma_p = \$2{,}000$. Ordinarily, the current procedure would be to approve the investment since $\mu_p > 0$. However, with the additional information available ($\sigma_p = \$2{,}000$) regarding the considerable risk of the investment, the executive can analyze the situation further. Using widely available tables for the normal distribution, he could note that the probability that NPV < 0, so that the investment won't pay, is 0.31. Furthermore, the probability is 0.16, 0.023, and 0.0013, respectively, that the investment will lose the present-worth equivalent of at least \$ 1,000, \$3,000, and \$5,000, respectively. Considering the financial status of the firm, the executive can use this and similar information to make his decision. Suppose, instead, that the executive is attempting to choose between this investment and a second investment with $\mu_p = \$500$ and $\sigma_p = \$500$. By conducting a similar analysis for the second investment, the executive can decide whether the greater expected earnings of the first investment justifies the greater risk. A useful technique for making this comparison is to superimpose the drawing of the probability distribution of NPV for the second investment upon the corresponding drawing for the first investment. This same approach generalizes to the comparison of more than two investments.

Let's see how Hillier obtained his probabilities:

$$P(\text{NPV}<0) = .31 \qquad P(\text{NPV}< -\$1{,}000) = .16$$
$$P(\text{NPV}< -\$3{,}000) = .023 \qquad P(\text{NPV}< -\$5{,}000) = .0013$$

by using the information $\mu_{\text{NPV}} = \$1{,}000$ and $\sigma_{\text{NPV}} = \$2{,}000$.

$$P(\text{NPV}<0) = P\left(Z \le \frac{0 - 1{,}000}{2{,}000}\right) = P(Z \le -.5) = P(Z \ge .5)$$
$$= .5 - .1915 = .31$$

$$P(\text{NPV}< -\$1{,}000) = \text{P}\left(\text{Z} \le \frac{-1{,}000 - 1{,}000}{2{,}000}\right) = P(Z \le -1.0) = P(Z \ge 1.0)$$
$$= .5 - .3413 = .16$$

$$P(\text{NPV}< -\$3{,}000) = \text{P}\left(\text{Z} \le \frac{-3{,}000 - 1{,}000}{2{,}000}\right) = P(Z \le -2.0) = P(Z \ge 2.0)$$
$$= .5 - .4772 = .023$$

$$P(\text{NPV}< -\$5{,}000) = \text{P}\left(\text{Z} \le \frac{-5{,}000 - 1{,}000}{2{,}000}\right) = P(Z \le -3.0) = P(Z \ge 3.0)$$
$$= .5 - .4987 = .0013$$

[12] How to calculate the NPV is explained in Appendix 4 (Chap. 21). In Hillier's example discussed below, he used P to represent NPV.

The above probability can be calculated directly by using the MINITAB program as indicated here:

```
MTB > SET INTO C1
DATA> 0-1000 -3000 -5000
DATA> END
MTB > CDF C1;
SUBC > NORMAL 1000 2000.
```

Cumulative Distribution Function

```
Normal with mean = 1000.00 and standard deviation = 2000.00
x P(X <= x)
0.0000 0.3085
-1-0E+03 0.1587
-3.0E+03 0.0228
-5.0E+03 0.0013
```

Application 7.4 Determination of Commercial Lending Rates.[13] The loan officers of a bank and the financial analysts of a firm seeking to borrow money consider the firm's total risks when analyzing the lending rate to the firm or—what is the same thing—the firm's cost of borrowing.

The lending rate is based partly on the risk-free rate (e.g., the federal government bond rate is free from default risk). First, we have to forecast the risk-free rate (R_f) for three economic conditions: boom, normal, and recession.

The second component of the lending rate is the risk premium (R_p). Risk premium is the bank's reward for taking risk. It can be calculated individually for each firm by examining the change in EBIT (earnings before interest and tax) under the three types of economic conditions. The EBIT is used as an indicator of the ability of the prospective borrower to repay borrowed funds.

Table 7.8 contains the information on R_f, EBIT, and R_p required for the analysis. It also shows the probability that each economic conditions will prevail (column B) and the probability of various levels of EBIT for the firm (column D).

A total of nine lending rates under the three different economic conditions are given in column F of Table 7.9. The probabilities of their occurrence under the different conditions are shown in column E.

Let us see how the numbers for columns E and F of Table 7.9 are calculated. During a period of normal economic conditions, the risk-free rate is at 10 %, as indicated in column A, but the risk premium can take on different values. There is a 40 % chance it will be 3.0 %, a 30 % chance it will be 5.0 %, and a 30 % chance it will be 8.0 %, as indicated in column D. We must multiply the probability for the risk-free rate by the conditional probability for the risk premium to get the probability of their occurring jointly (column E). The chance that the firm will receive a

[13] This application is similar to Applications 5.1 and Example 6.8. Note that the conditional probability used here is different from that used in Application 5.1 and Example 6.8.

Table 7.8 Worksheet for interest rate calculation

Economic condition	(A) R_f	(B) Marginal probability	(C) EBIT	(D) Conditional probability	(E) R_P
Boom	12.0 %	.25	$2.5m	.60	3 %
			1.5	.30	5
			.5	.10	8
Normal	10.0	.50	$2.5m	.40	3 %
			1.5	.30	5
			.5	.30	8
Poor (recession)	8.0	.25	$2.5m	.10	3 %
			1.5	.20	5
			.5	.70	8

Table 7.9 Worksheet for interest rate calculation

Economic condition	(A) R_f	(B) Marginal probability	(C) R_p	(D) Conditional probability	(E) Joint probability (B × D)	(F) Lending rate (A + C)
Boom	12 %	.25	3.0 %	.60	.150	15 %
			5.0	.30	.075	17
			8.0	.10	.025	20
Normal	10	.50	3.0 %	.40	.200	13 %
			5.0	.30	.150	15
			8.0	.30	.150	18
Poor (recession)	8	.25	3.0 %	.10	.025	11 %
			5.0	.20	.050	13
			8.0	.70	.175	16
					1.000	

13 % lending rate during normal economic conditions is 20 %; for a 15 % or 18 % rate, the probability is 15 % (see column F). This process also applies for the other two conditions (boom and recession).

For the problem set up in Table 7.9, the weighted-average lending rate is

$$\begin{aligned}\bar{R} &= (.150)(.150) + (.075)(.17) + (.025)(.20) + (.200)(.13) + \\ &\quad (.150)(.15) + (.150)(.180) + (.025)(.11) + (.050)(.13) + \\ &\quad (.175)(.16) \\ &= .1531 = 15.31\%\end{aligned}$$

with a standard deviation of

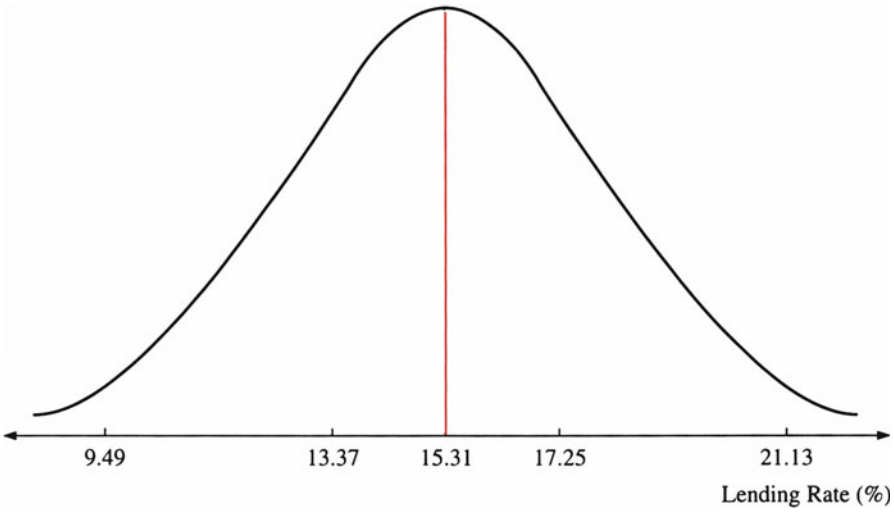


Fig. 7.16 Probability distribution for estimated lending rate

$$\begin{aligned}\sigma = &\Big[(.15)(.15-.153)^2+(.075)(.17-.153)^2+\\ &(.025)(.20-.153)^2+(.20)(.13-.153)^2+\\ &(.15)(.15-.153)^2+(.15)(.18-.153)^2+\\ &(.025)(.11-.153)^2+(.050)(.13-.153)^2+\\ &(.175)(.16-.153)^2\Big]^{1/2}\\ =&\,[.000001+.000021+.000055+.000107+\\ &.000001+.000109+.000046+.000027+.000008]^{1/2}\\ =&\,0.194\end{aligned}$$

We assume that the distribution of the lending rate is normal. Given the mean and standard deviation for such a distribution from Table 7.2, we see that 68.3 % of the observations of a normal distribution are within one standard deviation of the mean, 95.5 % are within two standard deviations, and 99.7 % are within three.

On the basis of the mean and standard deviation of the estimated lending rate, we can depict the expected lending rate and its standard deviation as shown in Fig. 7.16.

The percentages in Fig. 7.16, along with the mean and standard deviations, are an illustration of the normal distribution. The average lending rate is normally distributed with a mean of 15.31 % and a standard deviation of 1.94 %. This implies that almost all rates (99.7 %) will lie in the range of 9.49–21.13 %. We also know that 68.3 % of the rates will lie in the range of 13.37–17.25 %.

```
MTB> SET INTO C1
DATA>  .15 .17 .20 .13 .15 .18 .11 .13.16
DATA>  END
MTB >  SET INTO C2
DATA>  .150  .075  .025  .200  .150  .150  .025  .050  .175
DATA>  END
MTB" > LET C3=C1*C2
MTB >  SUM C3 INTO Kl
```

Column Sum

```
   Sum  Of C3 = 0.15300
MTB>   LET C4=C1-K1
MTB >  LET C5=C4**2
MTB >  LET C6=C5*C2
MTB >  SUM C6 INTO K2
```

Column Sum

```
   Sum  Of C6 = 0.00037600
MTB > LET K3=K2**.5
MTB > PRINT  C1-C6  K1-K3
```

Data Display

```
Kl    0.153000
K2    0.000376000
K3    0.0193907
```

Row	Cl	C2	C3	C4	C5	C6
1	0.15	0.150	0.02250	-0.003	0.000009	0.0000014
2	0.17	0.075	0.01275	0.017	0.000289	0.0000217
3	0.20	0.025	0.00500	0.047	0.002209	0.0000552
4	0.13	0.200	0.02600	-0.023	0.000529	0.0001058
5	0.15	0.150	0.02250	-0.003	0.000009	0.0000014
6	0.18	0.150	0.02700	0.027	0.000729	0.0001093
7	0.11	0.025	0.00275	-0.043	0.001849	0.0000462
8	0.13	0.050	0.00650	-0.023	0.000529	0.0000265
9	0.16	0.175	0.02800	0.007	0.000049	0.0000086

Fig. 7.17 MINITAB output for Application 7.4

The MINITAB output of the empirical calculation procedure for this problem is presented in Fig. 7.17.

7.7 Summary

In this chapter we introduced two of the most important continuous distributions, the normal and lognormal distributions. These distributions can be used to describe a wide variety of random variables in business, economics, and finance. In fact, even when our distribution is not normally distributed, it may be possible to transform our random variables (e.g., the logarithmic transformation discussed in

Sect. 7.4) so they are approximately normally distributed. This means that many of the analyses we will perform throughout the rest of the book will be based on the normal distribution. We illustrated this point by showing how the binomial and Poisson distributions could be approximated by using the normal distribution.

We also showed how the normal distribution could be applied to a variety of business problems, including EPS forecasting, CVP analysis, determining of commercial lending rates, and option pricing (see Appendix 2).

Questions and Problems

1. A study indicates that an assembly line task should take an average of 3.20 min to complete, with a standard deviation of 0.75 min. What is the probability that the task will take between 1.80 and 3.80 min to complete? Graph the area being determined by assuming that the completion time is normally distributed.
2. Indicate whether each of the following random variables is continuous or discrete.

   (a) The time it takes a mechanic to service a car.
   (b) The number of new housing starts in New Jersey this year.
   (c) The age of an applicant for an MBA program.
   (d) The sex of a new company chief executive officer.

3. The random variable Z is normally distributed with a mean of $\mu_z = 0$ and a standard deviation of $\sigma_z = 1$. Find the following probabilities:

   (a) $P(Z > 1.65)$
   (b) $P(Z > -2.38)$
   (c) $P(Z > 2.95)$
   (d) $P(Z < -1.37)$
   (e) $P(1.05 < Z < 2.82)$
   (f) $P(-2.43 < Z < 1.72)$

4. The random variable Z is normally distributed with $\mu_Z = 0$ and $\sigma_Z = 1$. Find the following values of b:

   (a) $P(Z < b) = .9280$
   (b) $P(Z > b) = .9949$
   (c) $P(Z > b) = .0074$
   (d) $P(Z < b) = .0130$
   (e) $P(-b < Z < b) = .5408$
   (f) $P(0 < Z < b) = .1844$

5. A random variable X is normally distributed with $\mu_X = 5$ and $\sigma_X = 2.7$. Find the following probabilities, using both manual calculations and the MINITAB program:

   (a) $P(X < 7.40)$
   (b) $P(X > -1.50)$

(c) $P(X>8.32)$
(d) $P(X<.95)$
(e) $P(2.35 < X < 7.05)$
(f) $P(-2.80 < X < 0)$

6. The random variable Y is normally distributed with $\mu_Y = 28.00$ and $\sigma_Y = 10$. Find the following values of b:

(a) $P(Y < b) = .8962$
(b) $P(Y > b) = .8106$
(c) $P(Y > b) = .0099$
(d) $P(Y < b) = .3409$
(e) $P(b < Y < 38) = .0227$
(f) $P(b < Y < 25) = .1148$

7. Suppose that X represents the number of cars arriving at a toll booth in 1 min. Further, suppose that X can assume the values 1, 2, 3, 4, and 5 and has the following distribution:

r	1	2	3	4	5
$P(X = r)$	.10	.20	.30	.25	.15

Calculate the expected value of this random variable and explain your result.

8. The following table gives the amount of time X, in seconds, by which an automated manufacturing process misses the designed completion time when performing a certain task. Negative values indicate early completion, and positive values late completion.

r	-1	0	1	2
$P(X = r)$	.1	.2	.3	.4

(a) Find the mean and the variance of X.
(b) On average, how do the completion times for this particular task compare with the designed completion times?

9. Find the probability density function of $Y = e^x$ when x is normally distributed with parameters μ and σ^2. The random variable Y is said to have a lognormal distribution (because log Y has a normal distribution) with parameters μ and σ^2.
10. Suppose that 35 % of the employees of the Harrison Company belong to unions. To determine union members' attitudes toward management, the company's personnel manager takes a random sample of 100 employees. The selection of a union member in this random sample is a "success." Calculate the probability that the number X of successes will be between 20 and 40, inclusive.
11. What is the probability that the number X of successes in the personnel manager's sample of 100 employees in question 10 will be 48 or more?
12. Suppose that a batch of $n = 80$ items is taken from a manufacturing process that produces a fraction $p = .16$ of defectives. What is the probability that this batch

will contain between 19 and 20 defectives? (Finding a defective is considered a success.)

13. The IQ scores of human beings are scaled to follow a normal distribution with mean 100 and standard deviation 16. If those with IQ scores higher than 154 are regarded as geniuses, how many geniuses are there among 20,000 children?
14. A college professor teaches corporate finance every semester. The tests for the course are standardized so that the test scores exhibit a normal distribution with a mean of 75 and a standard deviation of 12. The professor gives 15 % A, 25 % B, 30 % C, 20 % D, and 10 % F.

    (a) What letter grade will a student who scores 79 points on the test receive?
    (b) What letter grade will a student who scores 58 points receive?
    (c) How many points does a student need to score to get an A?
    (d) How many points does a student need to score to pass the course?

15. The manager in the local bank discovers that people come in to cash their paychecks on Friday. The amount of money withdrawn on Friday follows a normal distribution with 5 million as the mean and 1 million as the standard deviation. The bank manager wants to make sure that the amount of money in the bank can cover 99.9 % of the Friday withdrawals. What is the minimum amount of money he or she should plan to have on hand?
16. A local bakery found that it was throwing out too many cookies every night, so the manager conducted a study on the sales of cookies and found that on an ordinary day, the sales of cookies follow a normal distribution with 30 lb as the mean and 12 lb as the standard deviation. The manager then decided to prepare only 35 lb of cookies each day. What is the chance that the bakery will run out of cookies on a certain day?
17. A soft drink producer has just installed a new assembly line. The assembly line is adjusted to dispense an average of 12.05 oz of soda into the 12-oz soda can with a standard deviation of .02. What is the probability that certain cans will contain less than 12 oz of soda?
18. In question 17, what is the probability that 2 cans out of a six pack will contain less than 12 oz?
19. In question 17, the average amount of soda dispensed into the cans is adjustable. If we want to make sure that 99.9 % of the soda cans contain more than 12 oz, to what should we adjust the average?
20. A battery producer invents a new product. The life of the new battery is found to follow a normal distribution with a mean of 72 months and a standard deviation of 12 months. The producer guarantees that the new battery will last longer than 60 months or the full price will be refunded. Last year the producer sold one million batteries, how many refunds will be claimed?
21. A car manufacturer designs a fuel-efficient car for 1993. The company argues that the car can attain an average of 45 miles per gallon. The miles per gallon of the car follows a normal distribution with a standard deviation of 5. What is the probability that a certain car will reach 40 or more miles per gallon?

22. You work for a furniture factory that procures springs from an outside supplier. Every month, a truckload of springs comes in. From each shipment, you randomly inspect 400 springs. If there are 10 or more bad springs, then you send the shipment back. One day a shipment arrives that actually contains 3 % bad springs. What is the probability of your accepting this shipment?
23. A consumer rights organization wants to find out whether a local dairy farm actually puts 16 oz of milk into the container that is labeled 16 oz. Assume the milk put into the container by the local dairy farm follows a normal distribution with a mean of 16.05 and a standard deviation of .03.

    (a) What is the probability that a certain container contains more than 16 oz of milk?
    (b) The consumer rights organization bought 400 bottles of milk. What is the probability that among them, it found fewer than 12 bottles that do not contain enough milk?

24. A name-brand TV dinner boasts that its pot roast has no more than 120 calories per serving. Suppose 95 % of the servings of this product actually contain fewer than 120 calories. Find the probability that out of a random sample of 500 packs of pot roast, fewer than 10 packs (servings) actually contain more than 120 calories.
25. The Food and Drug Administration randomly tests 1,000 of a certain brand of cigarettes to see whether the nicotine content reaches a dangerous level. If 20 or more cigarettes contain more nicotine than a prespecified level, the production of the cigarettes is suspended. Assume that in this month, as a result of either machine failure or worker discontent, 3 % of the cigarettes contain more than the prespecified level. What is the probability that cigarette production will be suspended?
26. The light bulbs produced by Edison Lighting Corporation last an average of 300 h. The life of the light bulbs is believed to follow a normal distribution with a standard deviation of 10. A customer buys 2 dozen light bulbs during a sale. What is the probability that 1 light bulb used will last longer than 315 h?
27. The number of phone calls that reach 1–800 numbers in a certain time period follows a Poisson distribution. Assume that there are about 15,000 potential callers. Each caller has a probability of 0.001 of making such a phone call. What is the probability that we have less than 20 callers who make phone calls during this time period?
28. A camcorder is sold with a 1-year warranty. The probability that a camcorder is brought back for service under the warranty is 2 %. It costs the manufacturer $20 on average to repair a camcorder brought back under warranty. Last year, 5,000 camcorders were sold. What is the probability that fewer than 80 of them will be brought back to be repaired under warranty? How much should the company expect to spend living up to the warranty?
29. What are the advantages of using the lognormal distribution over using the normal distribution to describe stock prices?

30. Suppose that X is distributed as normal with a mean of 5 and a standard deviation of 2. Compute the standard normal values of X, given the following values of X:

    (a) 3
    (b) 2
    (c) 9
    (d) 11
    (e) 6
    (f) 10

31. Use the standard normal values you computed in question 30, and find the probability that Z is less than those values.
32. Calculate the area under the normal curve between the following:

    (a) $z = 0$ and $z = 2.0$
    (b) $z = -3.5$ and $z = -1$
    (c) $z = 1.2$ and $z = 3$
    (d) $z = -1.3$ and $z = 1.3$
    (e) $z = = 1$ and $z = 1$
    (f) $z = 3$ and $z = 4$

33. Find the value for z_0 for the following probabilities:

    (a) $P(z > z_0) = .10$
    (b) $P(z > z_0) = .75$
    (c) $P(-z_0 < z < z_0) = .95$
    (d) $P(z < z_0) = .95$
    (e) $P(-z_0 < z < z0) = .90$
    (f) $P(-z_0 < z < z_0) = 1.00$

34. Suppose that X is normally distributed with a mean of 5 and a standard deviation of 2. Find the following probabilities:

    (a) X is between 5 and 9.
    (b) X is between 0 and 8.
    (c) X is greater than 6.
    (d) X is less than 10.
    (e) X is between -1 and 3.

35. Briefly explain why it is useful for us to be able to approximate the binomial and Poisson distributions by using a normal distribution. Explain how we make this approximation.
36. Use the normal approximation to the binomial distribution with $n = 100$ and $p = .3$.

    (a) What is the probability that a value from the binomial distribution will have a value greater than 35?

(b) What is the probability that a value from the binomial distribution will have a value less than 20?
(c) What is the probability that a value from the binomial distribution will have a value between 15 and 45, inclusive?

37. Use the normal approximation to the binomial distribution with $n = 500$ and $p = .7$.

(a) What is the probability that a value from the binomial distribution will have a value greater than 325?
(b) What is the probability that a value from the binomial distribution will have a value less than 325?
(c) What is the probability that a value from the binomial distribution will have a value between 325 and 375, inclusive?

38. Use the normal approximation to the Poisson distribution with $\lambda = 75$.

(a) What is the probability that a value from the Poisson distribution will be greater than 50?
(b) What is the probability that a value from the Poisson distribution will be between 50 and 80, inclusive?
(c) What is the probability that a value from the Poisson distribution will be less than 60?

39. The time a customer waits for service at a bank is distributed normally with a mean of 4 min and a standard deviation of 1 min. Compute the probability that a customer must wait for:

(a) More than 10 min
(b) Less than 5 min
(c) Between 2 and 6 min
d. Between 3 and 9 min

40. The time it takes to get a car's oil changed at Speedy Lube is distributed normally with a mean of 12 min and a standard deviation of 2 min. Compute the probability that a customer will have her or his oil changed:

(a) In less than 9 min
(b) In between 9 and 15 min

41. A quality control manager has determined that the number of defective light bulbs in a case of 1,000 follows a normal distribution with a mean of 10 and a standard deviation of 3. Compute the probability that the number of defective light bulbs in a case is:

(a) Greater than 10
(b) Less than 9

42. A survey of recent masters of business administration (MBAs) reveals that their starting salaries follow a normal distribution with mean $48,000 and standard

deviation $9,000. Find the probability that a randomly selected MBA degree holder will begin his or her career earning:

(a) More than $50,000
(b) Less than $35,000

43. Use the Black–Scholes option pricing formula to compute the value of a call option, given the following information:

$S = \$55$	Price of stock
$X = \$50$	Exercise price
$r = .065$	Risk-free interest rate
$t = .5$	Time until the option expires, in years
$\sigma = .25$	Standard deviation of the stock's return

44. Answer question 43 again for an option with an exercise price of $55. How does the exercise value of the call option affect the option's price?
45. Answer question 43 again for an option with $t = .3$ years. How does the time until the option expires affect the value of the call option?
46. Answer question 43 again for an option whose stock price is $60. How does the price of the stock affect the value of the call option?
47. Answer question 43 again when $r = .10$. How does a change in the risk-free rate of interest affect the value of the call option?
48. Answer question 43 again when $\sigma^2 = .50$. How does a change in the variance of the stock's return affect the value of the call option?
49. Draw a standard normal probability function and show the area under the curve for

(a) Plus or minus one standard deviation from the mean
(b) Plus or minus two standard deviations from the mean
(c) Plus or minus three standard deviations from the mean

50. A company has a mean earnings per share (EPS) of $3.25 with a standard deviation of $1.21. Assume that the earnings are normally distributed. Compute the probability that EPS will be:

(a) Between $1.50 and $6.00
(b) Above $5.00

51. An investment analyst is following the stock of High Flyer Company. She believes that in any month, the stock has a 65 % chance of going up and a 35 % chance of going down. Using the binomial distribution, compute the probability that the stock goes up in 18 or more months during a 36-month period. Now use the normal approximation to the binomial distribution to recompute your answer. Compare the two results. Which method was easier to use?
52. A gas station finds that the mean number of people buying gas in any 30-min period is 16. Use the Poisson distribution to compute the probability that

between 25 and 35 people, inclusive, will buy gas in any 30-min period. Use the normal approximation to the Poisson to recalculate your result. Which method is easier to use?

53. Calculate e^y for the following values of y:

    (a) $y = 1$
    (b) $y = .5$
    (c) $y = -.5$
    (d) $y = -2.5$
    (e) $y = 3.1$
    (f) $y = -\text{I}$
    (g) $y = .05$
    (h) $y = .32$
    (i) $y = 6.1$
    (j) $y = -5.4$

54. Suppose $x = e^y$. Compute the value of y, given the following values of x:

    (a) $x = 2$
    (b) $x = 3$
    (c) $x \sim 1.5$
    (d) $x = .3$
    (e) $x = .5$
    (f) $x = .002$
    (g) $x = 10$
    (h) $x = 1$

55. Briefly explain what a cumulative distribution function is. Give some examples of occasions when the cumulative distribution function is useful.
56. A quality control manager has found that the mean number of ounces of cereal in a 16-oz box is 16 oz with a standard deviation of 2 oz. Calculate the probability that a randomly selected box of cereal will contain;

    (a) More than 16 oz of cereal
    (b) Less than 15 oz of cereal
    (c) Between 14 and 18 oz of cereal
    (d) Between 15 and 17 oz of cereal

57. An investment analyst calculates that the mean price of gold is $392 per ounce with a standard deviation of $12. Assume the price of gold follows a normal distribution. Compute the probability that the price of gold will be:

    (a) Greater than $400 an ounce
    (b) Less than $350 an ounce

58. You know that a certain stock's dividend yield has a mean of 6 % and a standard deviation of 2 %. Assume the dividends follow a normal distribution. Compute the probability that the dividend yield will be:

    (a) Less than 2 %

(b) Greater than 10 %
(c) Between 4 % and 8 %

59. Use the Black–Scholes option pricing formula to compute the value of a call option, given the following information:

$S = \$105$	Price of stock
$X = \$110$	Exercise price
$r = .055$	Risk-free interest rate
$t = .9$	Time until the option expires, in years
$\sigma = .45$	Standard deviation of the stock's return

60. The number of claims filed each week with Security Insurance Company has a mean of 700 and a standard deviation of 250. Calculate the probability that the number of claims this week will be:

(a) Greater than 1,000
(b) Less than 500
(c) Between 300 and 800
(d) Between 1,000 and 1,250

61. From past history, we know that 60 % of people audited by the IRS owe money to the government. If we take a random sample of 500 people who are being audited, what is the probability that between 280 and 320, inclusive, owe the IRS money?
62. The probability is .1 that a customer entering a food store will buy a can of coffee. If 1,000 customers enter the store, what is the minimum number of cans of coffee the store must have on hand to prevent the probability of running out of coffee from being higher than 5 %?
63. Determine the following probabilities. Assume that X follows a normal distribution:

(a) $P(80 \leq X \leq 95 \mid \mu = 92, \sigma = 10)$
(b) $P(X \geq 150 \mid \mu = 99, \sigma = 25)$

64. A public library has observed that the fine for overdue books is approximately normally distributed with a mean of \$2.72 and a standard deviation of \$.37.

(a) What is the probability that a fine will be greater than \$3?
(b) What is the probability that a fine will be less than \$2?

65. The breaking strength for paper bags used in a grocery store is approximately normally distributed with a mean of 15 lb and a standard deviation of 2 lb.

(a) What proportion of these bags has a breaking strength less than 10 lb?
(b) What proportion of the bags has a breaking strength greater than 17 lb?

66. A newspaper publisher has mean sales of 28,200 copies per day with a standard deviation of 3,100. If the publisher distributes 32,000 copies of the paper to the newsstands, what is the probability that at least 6,000 or more copies will go unsold?

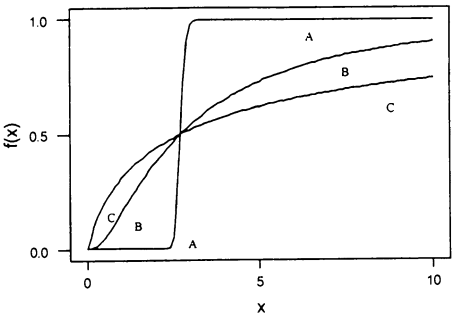
67. Value Line ranks 1,700 stocks according to their timeliness and riskiness. In other words, Value Line classifies these 1,700 stocks into five ranks (groups) on the basis of their return potential and the degree of riskiness as follows:

    These five groups are classified by assuming that they are normally distributed.

    (a) Find what percentage of the 1,700 stocks is classified in each group.
    (b) Calculate the mean and standard deviation of this ranking.

68. The following MINITAB output displays the cumulative distribution function curves of three normal distributions. Their mean and variance, respectively, are (0, .5), (0, 1), and (0, 2). Please compare the three cumulative distribution curves indicated in the figure.

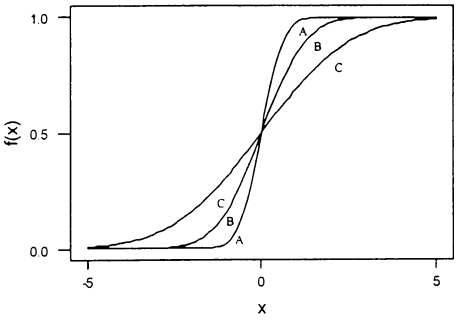
```
MTB > SET C1
DATA> −5;5/0.1
DATA> END
MTB > CDF C1 C2;
SUBC> NORMAL 0 0.5.
MTB > CDF C1 C3;
SUBC> NORMAL 0 1.
MTB > CDF C1 C4;
SUBC> NORMAL 0 2.
MTB > GPRO
* NOTE * Professional Graphics are enabled.
Standard Graphics are disabled.
Use the GSTD command to enable Standard Graphics.
MTB > Plot C2*C1 C3*C1 C4*C1;
SUBC> Connect,
SUBC> Type 1;
SUBC> Color 1;
SUBC> Size 1;
SUBC> Overlay;
SUBC> Axis 1;
SUBC> Label "X";
SUBC> Axis 2;
SUBC> Label "fx)".
```



MINITAB Output for Question 68

69. The following MINITAB output exhibits the cumulative distribution function curves of three lognormal distributions. Their mean and variance, respectively, are (1, .5), (1, 1), and (1, 2). Compare the three cumulative distribution curves indicated in the figure.

```
MTB > SET C1
DATA> 0:10/0.1
DATA> END
MTB > CDF C1 C2;
SUBC> LOGNORMAL 1 .05.
MTB > CDF C1 C3;
SUBC> LOGNORMAL 1 1.
MTB > CDF C1 C4;
SUBC> LOGNORMAL 1 2.
MTB > GPRO
* NOTE * Professional Graphics are enabled.
Standard Graphics are disabled.
Use the GSTD command to enable Standard Graphics.
MTB > Plot C2*C1 C3*C1 C4*C1;
SUBC> Connect;
SUBC> Type 1;
SUBC> Color 1;
SUBC> Size 1;
SUBC> Overlay;
SUBC> Axis 1;
SUBC> Label "x";
SUBC> Axis 2;
SUBC> Label "f(x)".
```



MINITAB Output for Question 69

Rank 1	Top 100
Rank 2	Next 300
Rank 3	Middle 900
Rank 4	Next 300
Rank 5	Bottom 100

70. The monthly earnings of financial analysts are normally distributed with a mean of $5,700. If only 6.68 % of the financial analysts have a monthly income of more than $6,140, what is the value of the standard deviation of the monthly earnings of the financial analysts?
71. Suppose 15 % of the parts produced by a machine are defective. Use a normal approximation to the binomial distribution. What is the probability that a sample of 50 parts contains:

    (a) Five or more defective parts?
    (b) Ten or fewer defective parts?
    (c) The expected number defective parts.
    (d) The standard deviation of the number of defective parts.

72. Suppose the grades of students were normally distributed with a mean of 73 and a standard deviation of 15.

    (a) If 10 % of her students failed the course and received Fs, what was the maximum score among those who received an F?
    (b) If 35 % of the students received grades of B or better (i.e., As and Bs), what is the minimum score of those who received a B?

73. A local bank has determined that the daily balances X of the checking accounts of its customers are lognormally distributed with an E(lnX)=$5.5 and Var(lnX) = 1.5.

    (a) What percentage of its customers has daily balances of more than $275?
    (b) What percentage of its customers has daily balances less than $243?
    (c) What percentage of its customers' balances is between $241 and $301.60?

Appendix 1: Mean and Variance for Continuous Random Variables

In this appendix, we will discuss areas under a continuous PDF and explore the variance for continuous random variables.

Areas Under Continuous Probability Density Function

In accordance with Eq. 6.2a, the *cumulative distribution function* (CDF) for a continuous variable X is given by

$$F(x_0) = \int_{-\infty}^{x_0} f(x)\, dx \tag{7.19}$$

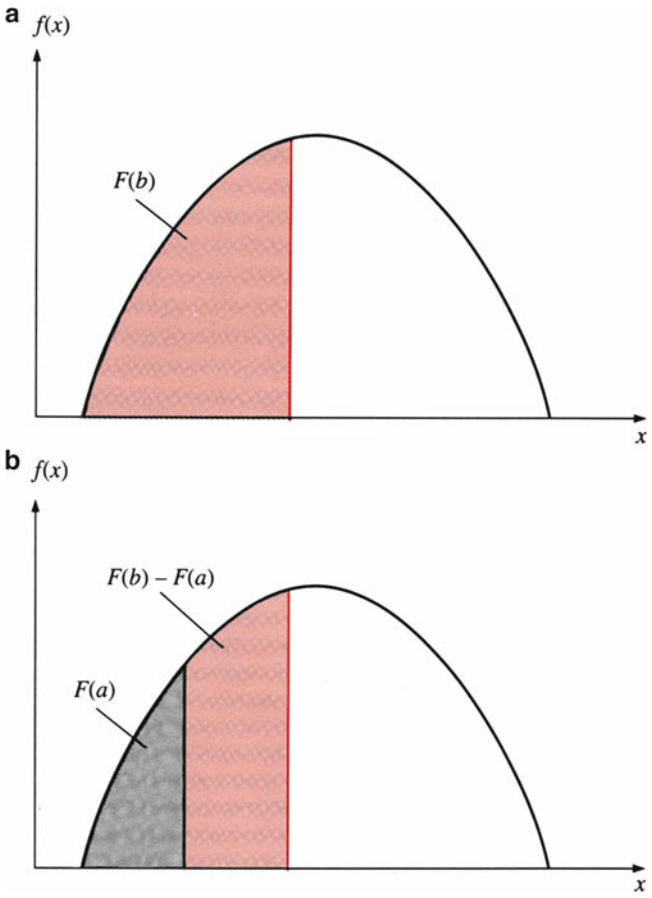


Fig. 7.18 Areas under continuous probability density function

where x_0 is any value that the random variable X can take. Eq. 7.19 implies that the area under curve $f(x)$ is to the left of x_0.

Using Eq. 7.19, we can calculate the probability that X lies between a and b for a continuous random variable:

$$P(a \leq X \leq b) = \int_a^b f(x)\, d(x)$$
$$= \int_{-\infty}^{b} f(x)\, dx - \int_{-\infty}^{a} f(x)\, dx = F(b) - F(a) \qquad (7.20)$$

For any $a \leq b$, $F(a) \leq F(b)$. Equation 7.20 is similar to Eq. 7.1 for a discrete variable case.

A diagrammatic representation of cumulative probability is given in Fig. 7.18. The total area under the curve$\int$(x) is 1. In integral calculus notation,

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{7.21}$$

Following Eqs. 7.19 and 7.20, the CDF of lognormal distribution can be defined as

$$\int_{a}^{\infty} f(x)dx \tag{7.22}$$

where f(x) is the probability density function (PDF) of a lognormal distribution. Its PDF is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \quad x > 0$$

In addition, it is well known that the PDF of a normal distributed variable y can be defined as

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right], \quad -\infty < y < \infty$$

By comparing the PDF of normal distribution and the PDF of lognormal distribution, we know that

$$f(x) = \frac{f(y)}{x} \tag{7.23}$$

In addition, it can be shown that[14]

$$dx = x\, dy \tag{7.24}$$

If we transform variable x in Eq. 7.22 into variable y, then the upper and lower limits of integration for a new variable are ∞ and ln a. Using this information, the CDF for lognormal distribution can be written in terms of the CDF for normal distribution as[15]

$$\int_{a}^{\infty} f(x)dx = \int_{\ln\, a}^{\infty} \left(\frac{f(y)}{x}\right) x\, dy = \int_{\ln(a)}^{\infty} f(y)dy \tag{7.25}$$

By substituting the PDF of normal distribution into the right-hand side of Eq. 7.25, it can be shown that

[14] From Eq. 7.5, we know that $x = e^y$. Then $dx = d(e^y) = e^y\, dy = xdy$.

[15] This relationship is obtained by substituting both Eqs. 7.23 and 7.24 into Eq. 7.22

$$\int_a^{\infty} f(x)dx = \int_{\ln(a)}^{\infty} f(y)dy = N(d) \tag{7.26}$$

where

$$d = \frac{\mu - \ln(a)}{\sigma}$$

This is the CDF of a lognormal distribution.

The value of $N(d)$ can be obtained from Table A3 of Appendix A as discussed in the text. Alternatively, the $N(d)$ can be approximated by the following formula:

$$N(d) \approx 1 - a_0 e^{-d^2/2}\left(a_1 t + a_2 t^2 + a_3 t^3\right) \tag{7.27}$$

where

$t = 1/(1+0.33267d)$
$a_o = 0.3989423$, $a_1 = 0.4361836$, $a_2 = -0.1201676$, $a_3 = 0.9372980$

Mean of Discrete and Continuous Random Variables

From Chap. 6, we know that the expected value of a discrete random variable can be defined as

$$\mu = E(X) = \sum_{i=1}^{N} x_i P(x_i) \tag{6.3}$$

The expected value of a continuous variable can be defined in a similar fashion. If X is a continuous random variable with probability density $f(x)$, its expected value is

$$E(X) = \int_{-\infty}^{\infty} xf(x)\, dx \tag{7.28}$$

We carry out the integration from $-\infty$ to $+\infty$ to make sure that all possible values of x are covered.

Let's look at an example of how Eq. 7.28 can be used to calculate the mean of a continuous variable. Suppose we let

$$\begin{aligned} f(x) &= 2(1-x), \quad 0 < x < 1 \\ &= 0 \qquad\qquad \text{otherwise} \end{aligned}$$

Substituting $f(x) = 2(1 - x)$ into Eq. 7.22, we obtain

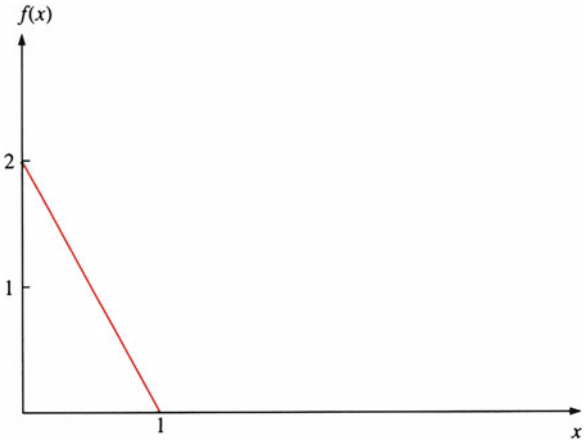


Fig. 7.19 Probability density function of $2(1 - x)$

$$\int_{-\infty}^{\infty} 2(1-x)\,dx = \int_{0}^{1} 2(1-x)\,dx$$
$$= 2x - x^2\Big|_0^1 = 1$$

Therefore, $f(x)$ is a PDF between $x \geq 0$ and $x \leq 1$. This PDF is shown in Fig. 7.19. The expected value of X in terms of $f(x) - 2(1 - x)$ can be calculated as

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{0}^{1} (x)2(1-x)\,dx$$
$$= x^2 - \frac{2}{3}x^3\Big|_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

From Eq. 7.6 in the text, the mean of a lognormal variable can be defined as

$$\int_{0}^{\infty} xf(x)dx = e^{\mu+1/2\sigma^2} \tag{7.29}$$

If the lower bound $\boldsymbol{a}$ is larger than 0, then the partial mean of x can be shown as[16]

$$\int_{0}^{\infty} xf(x)dx = \int_{\ln(a)}^{\infty} f(y)e^y dy = e^{\mu+\sigma^2/2}N(d) \tag{7.30}$$

where

[16] The first equality is obtained by using the technique to show Eq. 7.25. The second equality is obtained by substituting the PDF of normal distribution into $\int_{\ln(a)}^{\infty} f(y)e^y dy$ and do the appropriate manipulation.

$$d = \frac{\mu - \ln(a)}{\sigma} + \sigma$$

Since ***d*** is positive, therefore, $N(d) < 1$. This implies that the partial mean of a lognormal variable is the mean of x times an adjustment term, $N(d)$.

Variance for Discrete and Continuous Random Variables

For discrete variables, we calculate the variance by averaging the squares of all possible individual deviations about the mean. The variance is a measure of how spread out the observations are, and it indicates the general shape of a distribution. When all members of the population are obtainable and are used, we can define the variance of a discrete random variable as

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 P(x_i) \tag{6.4}$$

For a continuous random variable, the variance is defined as

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx \tag{7.31}$$

Equation 7.31 can be rewritten as[17]

$$\sigma^2 = E(X^2) - \mu^2 \tag{7.32}$$

where

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx \tag{7.33}$$

[17] From Eq. 7.31, we obtain

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f(x)\, dx \\
&= \int_{-\infty}^{\infty} (x^2) f(x)\, dx - 2\mu \int_{-\infty}^{\infty} x f(x)\, dx + \mu^2 \int_{-\infty}^{\infty} f(x)\, dx \\
&= \int_{-\infty}^{\infty} (x^2) f(x)\, dx - 2\mu^2 + \mu^2 \\
&= \int_{-\infty}^{\infty} (x^2) f(x)\, dx - \mu^2 = E(X^2) - \mu^2
\end{aligned}$$

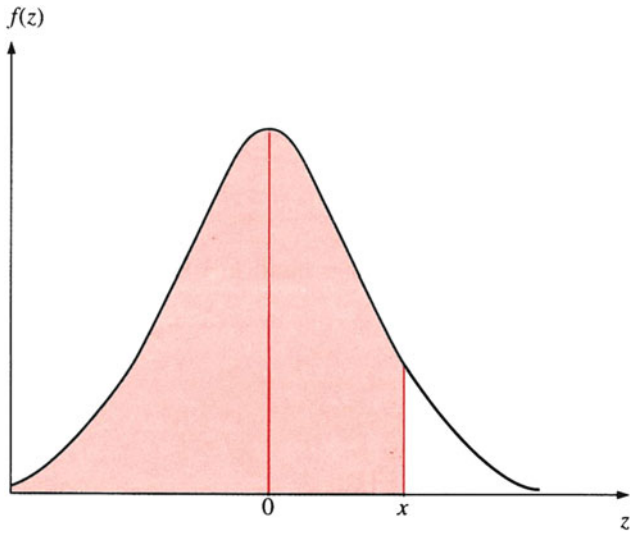


Fig. 7.20 $P(Z < x)$

Now let's see how we can use Eq. 7.32 to calculate the variance for a continuous random variable. For $f(x) = 2(1 - x)$, $E(X^2)$ can be calculated in accordance with Eq. 7.32 as

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \int_0^1 (x^2)2(1-x)\, dx$$
$$= \frac{2}{3}x^3 - \frac{2}{4}x^4\Big|_0^1 = \frac{2}{3} - \frac{2}{4} = \frac{1}{6}$$

Substituting $E(X^2) = \frac{1}{6}$ and $E(X) = \frac{1}{3}$ into Eq. 7.32, we obtain

$$\sigma^2 = E(X^2) - [E(X)]^2 = \tfrac{1}{6} - \tfrac{1}{9} = \tfrac{1}{18}$$

Appendix 2: Cumulative Normal Distribution Function and the Option Pricing Model

The cumulative normal density function tells us the probability that a random variable Z will be less than some value x. Note in Fig. 7.20 that $P(Z < x)$ is simply the area under the normal curve from $-\infty$ up to point x.

One of the many applications of the cumulative normal distribution function is in valuing stock options. Recall from Appendix 2 (Chap. 6) that a call option gives the

option holder the right to purchase, at a specified price known as the exercise price, a specified number of shares of stock during a given time period. A call option is a function of the following five variables:

1. Current price of the firm's common stock (S)
2. Exercise price of the option (X)
3. Term to maturity in years (T)
4. Variance of the stock's price (a)
5. Risk-free rate of interest (r)

From Appendix 2 (Chap. 6), the binomial *option pricing model* defined in Eq. 6.55 can be written as

$$\begin{aligned} C &= S\Big[\sum_{k=m}^{nT} \frac{nT!}{k!(nT-k)!} p'^{k}(1-p')^{nT-k}\Big] \\ &\quad - \frac{X}{(1+\frac{r}{n})^{Tn}}\Big[\sum_{k=m}^{nT} \frac{nT!}{k!(nT-k)!} p^{k}(1-p)^{nT-K}\Big] \\ &= SB(nT,p',m) - \frac{X}{(1+\frac{r}{n})^{Tn}} B(nT,p,m) \end{aligned}$$

where

n = number of periods per year of term to maturity (T)
T = term to maturity in years
m = minimum number of upward movements in stock price that is necessary for the option to terminate "in the money"

$p = \frac{R-d}{u-d}$ and $1-p = \frac{u-R}{u-d}$
where
$R = 1 + r = 1 +$ risk-free rate of return
$u = 1 +$ percentage of price increase
$d = 1 +$ percentage of price decrease

$$p' = \left(\frac{u}{R}\right)p$$

and

$$B(nT,p,m) = \sum_{k=m}^{nT} {}_{nT}C_k p^k (1-p)^{n-k}$$

By using Eq. 7.10 in the text and a form of the central limit theorem, when $n \to \infty$ the cumulative binomial density function can be approximated by the cumulative normal density function as

$$\begin{aligned} B_1(nT,p',m) &\cong N\left(Z_1, Z_1'\right) \\ B_2(nT,p,m) &\cong N\left(Z_2, Z_2'\right) \end{aligned}$$

where

$$Z_1 = \frac{m - nTp'}{\sqrt{nTp'(1-p')}}, \qquad Z_1' = \frac{nT - nTp'}{\sqrt{nTp'(1-p')}}$$
$$Z_2 = \frac{m - nTp}{\sqrt{nTp(1-p')}}, \qquad Z'_2 = \frac{nT - nTp}{\sqrt{nTp(1-p)}}$$

It can be shown that

$$\lim_{n\to\infty} \frac{1}{(1+\frac{r}{n})^{nT}} = e^{-rT} \quad \text{and} \quad \lim_{n\to\infty} Z_1' = \lim_{n\to\infty} Z_2' = \infty$$

Then Eq. 7.34 can be rewritten as

$$C = SN(Z_1) - Xe^{-rT}N(Z_2) \tag{7.34a}$$

Using the definition of m and property of lognormal distribution, it can be shown that Eq. 7.34a can become Eq. 7.35[18]:

$$C = SN(d_1) - Xe^{-rT}N(d_2) \tag{7.35}$$

where

C = price of the call option
S = current price of the stock
X = exercise price of the option
$e = 2.71828\ldots$
r = short-term interest rate (T-bill rate) = R_f
T = time to expiration of the option, in years
$N(d_i) = F_z(d_i)$ = value of the cumulative standard normal distribution ($i = 1, 2$)
σ^2 = variance of the stock rate of return

$$d_1 = \left[\ln(S/X) + \left(r + \frac{1}{2}\sigma^2\right)T\right] \Big/ \sigma\sqrt{T}$$

$$d_2 = d_1 - \sigma\sqrt{T}$$

If future stock price is constant over time, then $\sigma^2 = 0$. It can be shown that both $N(d_1)$ and $N(d_2)$ are equal to 1 and that Eq. 7.35 becomes

$$C = S - Xe^{-rT} \tag{7.36}$$

[18] See Rendleman, Jr R.J., Barter, B.J.: Two-state option pricing. J. Finance **34**, 1093–1010 (1979). The lognormal approach to derive the option pricing model will be shown in Appendix 3.

Alternatively, Eqs. 7.35 and 7.36 can be understood in terms of the following steps:

Step 1: The future price of the stock is constant over time.

Because a call option gives the option holder the right to purchase the stock at the exercise price X, the value of the option, C, is just the current price of the stock less the present value of the stock's purchase price. The concept of present value is discussed in Appendix 3 (Chap. 21) in detail. Mathematically, the value of the calf option is

$$C = S - \frac{X}{(1+r)^T} \tag{7.37}$$

Note that Eq. 7.37 assumes discrete compounding of interest, whereas Eq. 7.36 assumes continuous compounding of interest. To adjust Eq. 7.37 for continuous compounding, we substitute e^{-rT} for $1/(1 + \mathrm{r})^T$ to get

$$C = S - Xe^{-rT}$$

Step 2: Assume the price of the stock fluctuates over time (S_t).

In this case, we need to adjust Eq. 7.36 for the uncertainty associated with that fluctuation. We do this by using the cumulative normal distribution function. In deriving Eq. 7.35, we assume that S_t follows a lognormal distribution, as discussed in Sect. 7.4.[19]

The adjustment factors $N(d_1)$ and $N(d_2)$ in the Black–Scholes option valuation model are simply adjustments made to Eq. 7.36 to account for the uncertainty associated with the fluctuation of the price of the stock.

Equation 7.35 is a continuous option pricing model. Compare this to the binomial option pricing model given in Appendix 2 (Chap. 6), which is a discrete option pricing model. The adjustment factors $N(d_1)$ and $N(d_2)$ are cumulative normal density functions. The adjustment factors B_1 and B_2 are cumulative binomial probabilities.

We can use Eq. 7.35 to determine the theoretical value, as of November 29, 1991, of one of IBM's options with maturity on April 1992. In this case, we have $X = \$90$, $S = \$92.5$, $\sigma = 0.2194$, $r = 0.0435$, and $T = \frac{5}{12} = .42$ (in years).[20] Armed with this information, we can calculate the estimated d_1 and d_2:

[19] See Lee, C.F. et al.: Security Analysis and Portfolio Management, pp. 75–760. Scott, Foresman/Little, Brown, Glenview (1990)

[20] Values of $X = 90$, $S = 92.5$, and $r = .0435$ were obtained from Section C of the *Wall Street Journal* on December 2, 1991. And $\sigma = .2194$ is estimated in terms of monthly rates of return during the period January 1989 to November 1991.

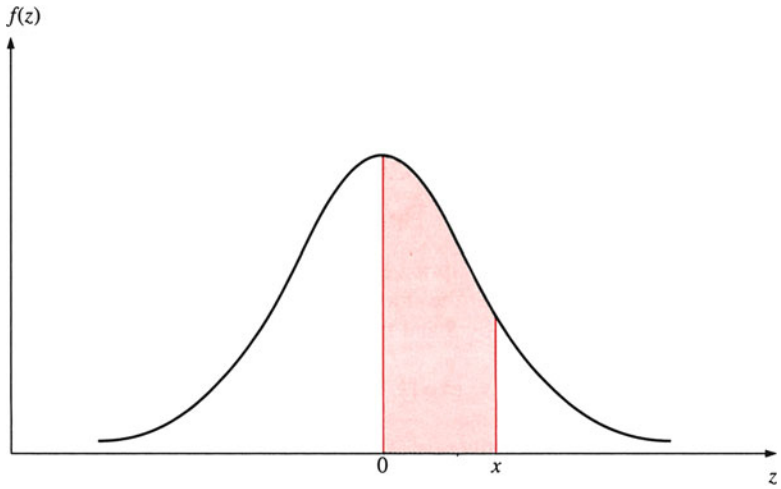


Fig. 7.21 $P(0 < Z < x)$

$$d_1 = \frac{\left\{\ln(92.5/90) + \left[(.0435) + \frac{1}{2}(.2194)^2\right](.42)\right\}}{(.2194)(.42)^{1/2}}$$
$$= .392$$
$$d_2 = d_1 - (0.2194)(0.42)^{1/2}$$
$$= .25$$

In Eq. 7.35, $N(d_1)$ and $N(d_2)$ are the probabilities that a random variable with a standard normal distribution takes on a value less than d_1 and a value less than d_2, respectively. The values for $N(d_1)$ and $N(d_2)$ can be found by using the tables in the back of the book for the standard normal distribution, which provide the probability that a variable Z is between 0 and x (see Fig.7.21).

To find the cumulative normal density function, we need to add the probability that Z is less than zero to the value given in the standard normal distribution table. Because the standard normal distribution is symmetric around zero, we know that the probability that Z is less than zero is .5, so

$$P(Z < x) = P(Z < 0) + P(0 < Z < x) = .5 + \text{value from table}$$

We can now compute the values of $N(d_1)$ and $N(d_2)$.

$$N(d_1) = P(Z < d_1) = P(Z < 0) + P(0 < Z < d_1)$$
$$= P(Z < .392) = .5 + .1517$$
$$= .6517$$
$$N(d_2) = P(Z < d_2) = P(Z < 0) + P(0 < Z < d_2)$$
$$= P(Z < .25) = .5 + .0987$$
$$= .5987$$

Then the theoretical value of the option is

$$\begin{aligned} C &= (92.5)(.6517) - [(90)(.5987)]/e^{(.0435)(0.42)}] \\ &= 60.282 - 53.883/1.0184 \\ &= \$7.373 \end{aligned}$$

and the actual price of the option on November 29, 1991, was $7.75.

Appendix 3: Lognormal Distribution Approach to Derive the Option Pricing Model

To derive the option pricing model in terms of lognormal distribution, we begin by assuming that the stock prices follow a lognormal distribution. Denote the current stock price by S and the stock price at the end of tth period by S_t.

Then $\frac{S_T}{S_{T-1}} = \exp[k_t]$ is a random variable with a lognormal distribution[21]

where K_t is the rate of return in tth period and is a random variable with normal distribution. Assume K_t has the same expected value μ_k and σ_k^2 for each t. Then $k_t + K_2 + \ldots + K_T$ is a normal random variable with expected value $T\mu_k$ and variance $T\sigma_k^2$.[22]

Following Eq. 7.6 in the text, we can define the expected value (mean) of $\frac{S_T}{S} = \exp[K_1 + K_2 + \ldots + K_T]$ as

$$E\left[\frac{S_T}{S}\right] = \exp\left[T\mu_k + \frac{T\sigma_k^2}{2}\right] \tag{7.38}$$

Under the assumption of a risk-neutral investor,[23] the expected return $E\left[\frac{S_T}{S}\right]$ is assumed to be $\exp[rT]$ (where r is the riskless rate of interest). In other words,

$$\mu_k = r - \sigma_k^2/2 \tag{7.39}$$

Following Appendix 2 (Chap. 6), the call option price C can be determined by discounting the expected value of the terminal option price by the riskless rate of interest:

$$C = \exp[-rT]E[Max(S_T - X, 0)], \tag{7.40}$$

where T is the time of expiration and X is the striking price.

[21] This is based upon the multiplicative property of lognormal distribution.

[22] This is based upon the additive property of normal distribution.

[23] The concept of risk-neutral investor will be discussed in Chap. 21 in detail.

$$\text{Note that } Max(S_T - X, 0) = \left(S\left(\frac{S_T}{S} - \frac{X}{S}\right)\right) \quad \text{for } \frac{S_T}{S} > \frac{X}{S}$$
$$= 0 \quad \text{for } \frac{S_T}{S} < \frac{X}{S} \tag{7.41}$$

Let $x = \frac{S_T}{S}$ has a lognormal distribution. Then

$$\begin{aligned} C &= \exp[-rT]\, E[Max(S_T - X)] \\ &= \exp[-rT] \int_{\frac{X}{S}}^{\infty} S\left[x - \frac{X}{S}\right] g(x)dx \\ &= \exp[-rT]S \int_{\frac{X}{S}}^{\infty} xg(x)dx - \exp[-rT]S.\frac{X}{S}\int_{\frac{X}{S}}^{\infty} g(x)dx \end{aligned}$$

where $g(x)$ is the probability density function of $X_T = X_T = \frac{S_T}{S}$.

Substituting $\mu = r - \sigma^2/2$ and $a = \frac{X}{S}$ into Eqs. 7.26 and 7.30, we obtain

$$\int_{\frac{X}{S}}^{\infty} xg(x)dx = e^r N(d_1)$$

where

$$d_1 = \frac{r - (1/2)\sigma^2 - \ln\frac{X}{S}}{\sigma} + \sigma \tag{7.43}$$

$$\int_{\frac{X}{S}}^{\infty} g(x)dx = N(d_2)$$

$$d_2 = \frac{r - (1/2)\sigma^2 - \ln\frac{X}{S}}{\sigma} \tag{7.44}$$

Substituting Eqs. 7.43 and 7.44 into Eq. 7.42, we obtain

$$C = SN(d_1) - X\exp[-rT]N(d_2), \tag{7.45}$$

$$d_1 = \frac{\ln\left(\frac{S}{X}\right) + \left(r + \frac{1}{2}\sigma_k^2\right)T}{\sigma_k\sqrt{T}}$$

$$d_2 = \frac{\ln\left(\frac{S}{X}\right) + \left(r - \frac{1}{2}\sigma_k^2\right)T}{\sigma_k\sqrt{T}} = d_1 - \sigma_k\sqrt{T}$$

This is Eq. 7.35 defined in Appendix 2.

	A	B
1	The Black-Scholes Pricing Formula Calculation	
2		
3	S(current stock price)=	42
4	X(exercise price of option)=	40
5	r(risk-free interest rate)=	0.1
6	σ(volatility of stock)=	0.2
7	T-t(expiraiion date of option · current time)=	0.5
8	d1 =	=(LN(B3/B4)+(B5+B6^2/2)*(B7))/(B6*SQRT(B7))
9	d2=	=(LN(B3/B4)+(B5-B6^2/2)*(B7))/(B6*SQRT(B7))
10		
11	c(value of European call option to buy one share)=	=B3*NORMSDIST(B8)-B4*EXP(–B5*B7)*NORMSDIST(B9)
12	p(value of European put option to sell one share}=	=B4*EXP(-B5*B7)*NORMSDIST(-B9)-B3*NORMSDIST(-B9)

Fig. 7.22 The excel calculations of The Black-Scholes Pricing Formula

	A	B
1	The Black-Scholes Pricing Formula Calculation	
2		
3	S(current stock price)=	42
4	X(exercise price of option)=	40
5	r(risk-free interest rate)=	0.1
6	σ(volatility of stock)=	0.2
7	T-t(expiration date of option -current time)=	0.5
8	d1 =	0.7693
9	d2=	0.6278
10		
11	c(value of European call option to buy one share)=	4.76
12	p(value of European put option to sell one share)=	0.81

Fig. 7.23 The Black-Scholes Pricing Formula calculation

In Appendix 2 (Chap. 6), we have defined a *put option* as a contract conveying the right to sell a designated security at a stipulated price. It can be shown that the relationship between a call option (C) and a put option (P) can be defined as[24]

[24] The relationship is known as *put-call parity*. See Hall, J.C.: Introduction to Futures and Option Markets Prentice-Hall, New Jersey (1995)

$$C + Xe^{-rT} = P + S \tag{7.46a}$$

Substituting Eq. 7.45 into 7.46a, we obtain the put option formula as

$$P = Xe^{-rT}N(-d_2) - SN(-d_1) \tag{7.46b}$$

where S, C, r, T, d_1, and d_2 are identical to those defined in the call option model.

Example 7.8 Excel Program for Calculating Black–Scholes Call and Put Option Models. Assume $S = \$42$, $X = 40$, $r = 0.1$, $\sigma = 0.2$, and $T - t = 0.5$

Using Eqs. 7.45 and 7.46a, we can write the Excel program for calculating the call and put option program as presented in Fig. 7.22. The results are presented in Fig. 7.23. From Fig. 7.23, we obtain $C = \$4.76$ and $P = \$0.81$.

Chapter 8
Sampling and Sampling Distributions

Chapter Outline

Key Terms

Census	Systematic error
Population	Sampling costs
Sample	Cost–benefit analysis of sampling
Simple random sampling	Sampling distribution
Random sample	Central limit theorem
Sampling errors	Finite population multiplier
Random errors	Sample proportion
Nonsampling errors	Confidence interval

8.1 Introduction

In this chapter, we take an in-depth look at the operational end of statistical analysis. Statistical analysis primarily involves selecting parts of populations (known as samples) and analyzing them in order to make inferences about the populations. Inferences made about a population by using sample data are widespread in business,

C.-F. Lee et al., *Statistics for Business and Financial Economics,*
DOI 10.1007/978-1-4614-5897-5_8,

economics, and finance. For example, the A. C. Nielsen Company infers the number of people who watch each television show on the basis of a sample of TV viewers. The use of political polls to project election winners is another example of statistical inference. And when you fill out a warranty card on an appliance you have bought, you are often asked to provide information about yourself that the warrantor compiles (and probably sells to someone who will later try to convince you to buy a magazine subscription). These data are also sample data.

First, sampling from a population is discussed. Second, we explore the issue of sampling costs versus sampling errors. Next, sampling distributions for sample means and sample proportions are illustrated. Then one of the most important principles in statistics, the central limit theorem, and confidence intervals are discussed in detail. Finally, an accounting application illustrates how sampling and sampling distributions can be used in auditing. The sampling distribution concept also is used to do patient waiting-time analysis.

8.2 Sampling from a Population

In previous chapters, we have discussed many different topics in statistics. Among these are distributions, probabilities, measurements of dispersion and symmetry, and data collection and analysis. The topic most closely related to sampling is data collection, organization, and presentation, which we discussed in Chap. 2. Either a census or a sampling survey approach can be used in data collection. A *census* is a survey that attempts to include every element in the universe, or population, in which we are interested. Sampling is used to count or measure only a subset of the population; these collected data are called sample data. In this book, we will return again and again to problems whose solutions depend on making inferences about a population from a sample.

The management, analysis, and interpretation of data are the foundation of statistics. In order to make full use of the information that data can yield, the statistician must start with clear objectives and follow well-defined pathways to a desired result. Along these pathways, there are points where the analyst must make decisions on the basis of an evaluation of costs and benefits. Key considerations include how much information is appropriate, how specific this information should be, and whether statistical inferences drawn from the data are analytically sound.

Now let us formally define the terms *population* and *sample*. A *population* consists of all members, objects, or observations that fall into a certain category. A *sample* is a subset of the members, objects, or observations in a given population.

In analyzing the characteristics of a population, a researcher can analyze the entire population or draw conclusions about the entire population on the basis of a random sample selected from the population. Using sampling to determine the true characteristics of a population offers several advantages:

1. The cost is less.
2. The data are more manageable.

3. It is less time-consuming.
4. Sample observation can be more accurate.
5. It makes analysis possible even when not all population elements are accessible.

As in other areas of statistics, the goal of the researcher is to choose methods that will lead to informative and useful results. These issues are discussed in the following section.

Sampling enables a statistician (or researcher) to make inferences about a given population from a more manageable segment of the population. It follows that the sample must be chosen in a manner that will ensure that it represents the original population. Two kinds of errors can arise in a sampling experiment: sampling errors and nonsampling errors. Before we can examine samples, we must thoroughly understand these two types of errors.

8.2.1 Sampling Error and Nonsampling Error

Sampling errors are errors that result from the chance selection of sampling units. They occur only when a sample, rather than the entire population, is observed. They are *random errors* (or chance errors), as discussed in Chap. 2. For example, if the sample from a given population had a mean of .6 and the true population mean was .5, there is an average sampling error of .1. Sampling error can be reduced by taking more observations, and it can be eliminated by taking all observations. Sampling error can usually be analyzed by first identifying the source of the error and then making the needed inferences. The relationship between sampling cost and sampling error is discussed in the next section of this chapter.

Nonsampling errors are errors that result from inaccurate measurement of the data or improper selection of sample observations. For example, if you measure flour with a cup that holds 15 oz rather than 16, the bread you make will contain less flour than the recipe intended. This kind of error is not related to the number of observations but rather is due to the inaccurate measurement of data. If a given section of the population has an unduly low or an unduly high chance of being selected for a sample, then sampling data can result in systematic, rather than random, sampling error. Other examples of nonsampling error include faulty questions and choosing observations that do not pertain to the population being examined. Nonsampling error is *systematic error* (or bias), as discussed in Chap. 2.

Unlike sampling error, nonsampling error cannot be reduced by increasing the sample size. (This issue will be discussed further in Chap. 20.) Although it is possible to minimize this type of error by carefully specifying the criteria by which observations are selected or measured, nonsampling error persists to a certain extent in almost all cases. The more complicated the data set, the greater the chance that nonsampling error will creep in.

8.2.2 Selection of a Random Sample

For a sample to be drawn from a population representatively, each member, object, or observation must have an independent and equal chance of being selected for the sample. Suppose a sample of n elements must be selected from a population of N elements. A *simple random sampling* procedure is one in which every possible combination $\left[\binom{N}{n}\right]$ of n elements in the population has an equal probability of being selected. The n elements obtained from simple random sampling constitute a simple random sample or *random sample*. Random selection is the key to this process; it significantly reduces nonsampling errors due to improper selection of sample observations.

There are two useful methods for carrying out simple random sampling: drawing chips from a box and using random-number tables.

8.2.2.1 Drawing from a Box

If we want to draw, with replacement, a simple random sample of five students from a business statistics class made up of 80 students, we assign the numbers 1–80 to the students and place these numbers on physically similar balls, slips of paper, or poker chips. We then put all the balls (or whatever) in a box, shake the box to mix them thoroughly, and proceed to draw the sample. The first ball is drawn, and we record the number written on it. We then replace the ball and shake the box again, draw the second ball, and record the result. We repeat the process until we have drawn five distinct numbers. The students corresponding to these five numbers constitute the required simple random sample.

8.2.2.2 Using a Random-Number Table

If the population size is large, the method just described becomes unwieldy and time-consuming. Furthermore, it may introduce biases if the balls are not thoroughly mixed. Using such random-number tables as Table 8.13 in Appendix 1 to draw random samples is much easier. A table of random digits is simply a table of digits generated by a random process. The application of random-number tables to draw random samples will be thoroughly discussed in Chap. 20.

MINITAB, SAS, and other computer programs can be used to generate random numbers. Both Tables 8.1 and 8.2 are generated from MINITAB. Table 8.1 contains the instructions for generating 200 random numbers between 1 and 1,000 in terms of a uniform distribution. Table 8.2 contains the instructions for generating 200 numbers between 0 and 1 in terms of a uniform distribution.

Table 8.1 Generating 200 random numbers between 1 and 1,000 in terms of a uniform distribution by using MINITAB

```
MTB > RANDOM 200 VALUES INTO C1; SUBC > UNIFORM A=1 AND B=1000.
MTB > PRINT C1
```

Data display

C1

941.986	39.449	346.863	383.933	11.275	276.998	536.211	953.945
855.417	724.206	820.295	626.853	975.945	566.319	843.182	235.036
367.646	630.667	664.626	381.481	131.648	275.402	784.816	290.423
750.941	816.607	442.353	620.808	621.520	227.820	522.829	720.876
926.772	309.486	931.945	282.630	465.286	73.395	768.555	869.484
505.738	535.606	142.547	292.865	281.997	656.920	983.971	286.044
205.122	633.400	693.314	470.996	398.395	268.005	954.336	65.553
643.058	633.837	584.367	306.635	329.898	756.054	737.467	65.848
508.341	519.965	326.375	610.217	634.219	533.356	987.627	995.352
99.847	160.561	454.765	43.812	144.597	440.315	657.047	381.957
245.586	117.927	417.549	518.071	570.668	494.793	67.074	405.027
597.375	332.791	834.783	476.291	484.948	512.582	357.934	617.527
298.144	682.348	138.652	41.604	672.946	603.005	316.132	533.274
22.857	414.416	579.174	45.299	866.907	540.638	569.469	91.323
398.065	453.727	618.530	930.257	850.171	11.154	164.686	994.463
796.081	588.330	687.029	530.520	937.161	730.301	512.738	929.849
501.783	527.130	322.678	716.984	830.000	445.996	717.680	651.760
254.035	832.549	680.402	931.214	847.747	801.548	273.196	949.470
165.742	731.310	422.866	585.386	533.260	89.009	135.737	489.258
16.525	795.346	976.937	338.351	631.684	862.639	176.294	524.535
919.020	502.837	176.764	249.829	448.567	444.515	3.789	490.098
366.807	669.368	267.488	806.726	577.593	286.242	163.099	464.508
643.956	759.218	299.569	827.572	8.952	220.808	710.697	510.594
626.208	67.765	885.748	237.092	605.323	581.556	164.924	890.085
943.773	777.026	341.626	717.893	963.489	200.278	824.881	155.359

Table 8.2 Generating 200 random numbers between 0 and 1 in terms of a uniform distribution by using MINITAB

```
MTB > RANDOM 200 VALUES INTO C1;
SUBC> UNIFORM A=0 B=1.
MTB > PRINT C1
```

Data display

C1

0.527281	0.896209	0.191064	0.242525	0.640217	0.273716	0.477429
0.108435	0.523726	0.064086	0.061118	0.471682	0.758676	0.116652
0.465407	0.650081	0.590382	0.234776	0.937043	0.032366	0.989526
0.776318	0.138995	0.486266	0.562677	0.236521	0.802187	0.670882
0.906173	0.532067	0.679966	0.722118	0.497111	0.205060	0.649055
0.279895	0.715202	0.265671	0.300735	0.325926	0.089223	0.892219
0.866023	0.800776	0.361845	0.707411	0.818573	0.468981	0.466556
0.849510	0.999064	0.000949	0.010451	0.557175	0.604829	0.740422
0.356630	0.486719	0.888459	0.126227	0.899954	0.480545	0.909149
0.237077	0.629911	0.005500	0.180522	0.428164	0.574341	0.979899
0.558693	0.528308	0.217774	0.207976	0.214314	0.566393	0.488122
0.140500	0.032354	0.495856	0.405480	0.180323	0.283365	0.885972
0.986797	0.251589	0.909978	0.048684	0.500138	0.851371	0.272390
0.569729	0.662327	0.302480	0.696580	0.712819	0.873678	0.426194
0.669012	0.665603	0.083602	0.353657	0.021568	0.096197	0.380816
0.694705	0.950474	0.988990	0.901910	0 782861	0.212954	0.228509
0.152027	0.215250	0.345366	0.900083	0.331403	0.406190	0.246581
0.776042	0.761511	0.373202	0.970338	0.956223	0.004570	0.474190
0.070336	0.131521	0.770161	0.314552	0.341818	0.439235	0.202702
0.542948	0.303372	0.378959	0.643534	0.754506	0.262736	0.757776
0.461365	0.220188	0.310323	0.133493	0.044820	0.661080	0.084238
0.724678	0.550863	0.786361	0.650527	0.256236	0.502809	0.813686
0.342409	0.709292	0.831712	0.690744	0.902243	0.444112	0.454857
0.236737	0.769985	0.664830	0.985107	0.922530	0.861869	0.694708
0.664864	0.508230	0.683081	0.030711	0.118390	0.256923	0.118162
0.646111	0.016233	0.943348	0.742900	0.509903	0.245517	0.376231
0.001455	0.791757	0.679842	0.653049	0.092461	0.721270	0.142533
0.865960	0.907877	0.570270	0.657016	0.477661	0.868309	0.726909
0.578034	0.085909	0.998574	0.330310			

Whichever sampling method is used, the analyst must be sure the population under consideration is appropriate for the analysis; taking this precaution largely eliminates another kind of nonsampling error.

8.3 Sampling Cost Versus Sampling Error

This section deals primarily with the costs associated with selecting a sample and with how those *sampling costs* can affect sampling errors. This type of analysis is often referred to as *cost–benefit analysis* of sampling. Cost–benefit analysis in this context involves comparing the benefits of sampling with its disadvantages (costs). The underlying need for the information is the gauge by which incurred costs and allowable error are measured. This issue will be explored further in Chap. 20.

The aim of drawing a random sample from a population is to measure indirectly population attributes such as mean and variance without having to include all possible data. We have all heard the saying "time is money." This is the heart of this issue. It takes people (and usually machines as well) to work through detailed analyses, and neither of these resources is free. A researcher must pay employees to collect the data and enter them into a computer; it also costs a lot to buy, use, and run the computer. The computer costs are numerous: hardware, software, electricity, paper, maintenance, operators, storage, and so on. If the computer is rented, these costs are included in the rental fee. Either way, the more data collected and analyzed, the higher the costs of the study. It is obvious that the statistician faces a crucial question: How much data are actually necessary?

The Gallup Organization and National Opinion Research Center used a sampling survey approach in their poll to obtain Americans' views on their work ethic in a timely manner. The results of the poll were published in the *Wall Street Journal* (February 13, 1992, p. Bl).

The three questions asked and the results are presented here.

Would you welcome or not welcome less emphasis on working hard?

Would not	67 %
Would	30 %

Are you satisfied or dissatisfied with Americans' willingness to work hard to better themselves?

Dissatisfied	45 %
Satisfied	52 %

Would you strongly agree, agree, disagree, or strongly disagree with the following:

"I am willing to work harder than I have to in order to help this organization succeed."

Strongly disagree	1 %
Disagree	9 %
Agree	52 %
Strongly agree	38 %

Note that the percentages for questions 1 and 2 add up to only 97 % because the category of "no response" was omitted during the poll.

8.3.1 Sampling Size and Accuracy

If an entire population is used as a sample for analysis, then such numerical characteristics as the mean and variance of the sample are identical to those of the population. However, suppose the population is very large – say, 10 million units. To collect all the observations and analyze them would be a ponderous task. Fortunately, if only some of the members are chosen at random and analyzed, the population mean and variance can be estimated with some precision from the sample. Even though it is possible to estimate the population parameters by analyzing a random sample of the observations, the results are only estimates. In general, the fewer the observations used in the sample, the larger the sample error. Significantly, sample error is not necessarily a linear function; that is, there is not necessarily an equal trade-off between additional data and greater accuracy. A relatively small sample of the entire population may yield estimates close to the true population values. However, it generally takes increasing amounts of data to make sample estimates closer to the true population value. Consequently, to have the sample estimates equal the population parameters is very expensive. The following two applications may shed some light on the problem of whether large or small samples should be used in the real world.

Application 8.1 A Case for a Large Sample. Suppose a pharmaceutical firm wishes to test a new shampoo formulated to help control dandruff for an acceptable amount of a certain active ingredient. If there is not enough of the active ingredient, the shampoo is not effective, yet if too much of the active ingredient is present, the shampoo may cause harmful side effects, including hair loss. Although there is a great need for accuracy, it is not economically feasible to test an entire batch of the shampoo. A sample can be used to test the content of the shampoo and to conduct related analyses and make inferences. In this case, it is particularly important to work with a large sample in order to reduce sampling errors because hair loss among users would be an intolerable outcome.

Application 8.2 A Case for a Small Sample. Suppose a company manufactures a crude grade of cement mix to be used as a foundation for sidewalks. The company wishes to check that a certain amount of small stones is included in each 50-lb bag. ALL components of the cement mix are equally valuable, so the only reason for conducting this test is to ensure the most durable mixture possible. A few stones more or less in a bag of cement mix will negligibly affect the performance of the

cement mix. Therefore, the producer will want only a small sample of the total number of cement mix bags to be examined to ensure that the stone content of each is within a certain range. Here, the underlying need for accuracy is small, so the company can save money by examining its product infrequently. When extremely high accuracy is not required, cost considerations and time constraints usually hold down sample size.

8.3.2 Time Constraints

If there is a deadline to be met, that in itself may limit the number of observations that are analyzed in a given study. For example, if we want to know the monthly inflation rate of the United States of America for an economic policy decision, we can use only a small number of sample data to calculate the monthly inflation rate in time. (How to use a price index to calculate the inflation rate will be discussed in Chap. 19.)

In general, sampling cost and sampling error are traded off according to the needs of the situation. The greater the accuracy required, the lower the allowable sampling error and the higher the cost of analysis. The issue of trade-offs between sampling cost and sampling error will be analyzed in more detail in Chap. 20.

In addition to the examples discussed so far, we turn to the real-world example of a telephone sampling survey used to find out about the different opinions among Americans and Japanese regarding their trade relationship.

Infoplan/Yankelovich International polled 500 Japanese adults via telephone on January 28 and 29 of 1992; 1,000 American adults were surveyed via phone by Yankelovich Clancy Shulman on January 30. The results of the TIME/CNN sponsored poll that posed questions about how Japanese and Americans feel about each other were published in the February 10, 1992, issue of *TIME*. Sampling errors are plus or minus 4.5 % for the survey of Japanese and 3 % for the survey of Americans. Responses of "not sure" were omitted.

Here are the results of 1 of the 5 questions in the *TIME* article.

Which is the main reason for the large trade imbalance between the United States and Japan?

1. Sixty-six percent of the Americans and 33 % of the Japanese surveyed responded that Japan unfairly keeps American products out of the country.
2. "American products are not as good as Japanese products," according to 22 % of the American respondents and 44 % of the Japanese.

8.4 Sampling Distribution of the Sample Mean

In previous chapters, we examined population distributions. Now we will examine *sampling distributions* of the sample mean. The sampling distribution is derived from a set of values taken at random from the population. In short, the population

Table 8.3 Work experience for six secretaries in Francis Engineering, Inc

Secretary	Mary	Gerry	Alice	Debbie	Elizabeth	Kimberly
Years of experience	1	2	3	4	5	6

distribution represents the distribution of the members of a population, whereas the sample distribution represents the distribution of a sample statistic for certain randomly chosen members of a population.

8.4.1 All Possible Random Samples and Their Mean

The following example shows how to calculate all possible random samples and their mean.

Example 8.1 Sampling Distribution: Three Cases. Consider the data in Table 8.3, which consist of the numbers of years of work experience for six secretaries in Francis Engineering, Inc.

The mean of this population is

$$\mu = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$$

A sample mean as indicated in Eq. 8.1 can be used to estimate this population mean:

$$\overline{X} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} X_i \tag{8.1}$$

where $X_1, X_2, \ldots, X_n$ denote the sample observations.

This example will show how the distribution of the sample mean can be affected by the sample size n as in the cases here.

Case 1: $n = 2$

Table 8.4 shows the possible values for a sample consisting of two observations from the above-mentioned population. Table 8.4 indicates that there are 15 possible samples. Because all are equally likely to be selected, the probability that any specific sample will be selected is 1/15. Using this information, we can summarize the probability distribution associated with $\overline{X}$ indicated in Table 8.4 as shown in Table 8.5. Now look at Fig. 8.1. Part (a) is the population distribution of work experience for six secretaries, which is a uniform distribution. Part (b) shows the sampling distribution of the mean for a sample size of 2; the information on which it is based is taken from Table 8.5. Note the difference between these probability distributions. That in part (b) looks more like the bell shape of the normal distribution. The numbers in the population range from 1 to 6 and the sample means have a more narrow range – from 1.5 to 5.5.

Table 8.4 Possible samples and sample means ($n = 2$)

Sample	Sample mean	Sample	Sample mean
1, 2	1.5	2, 6	4
1, 3	2	3, 4	3.5
1, 4	2.5	3, 5	4
1, 5	3	3, 6	4.5
1, 6	3.5	4, 5	4.5
2, 3	2.5	4, 6	5
2, 4	3	5, 6	5.5
2, 5	3.5		

Table 8.5 Probability function of $\overline{X}$ for $n = 2$

$\overline{X}$	$P(\overline{X})$
1.5	1/15
2	1/15
2.5	2/15
3	2/15
3.5	3/15
4	2/15
4.5	2/15
5	1/15
5.5	1/15

Case 2: $n = 3$

If the sample size is increased from 2 to 3, then the sample means and probabilities are as shown in Tables 8.6 and 8.7. Comparing Table 8.7 with Table 8.5 reveals that the range of possible values of the sample mean has been reduced from (5.5–1.5) to (5–2) – that is, from 4 to 3. Note that Fig. 8.2 looks more like a bell-shaped normal distribution than Fig. 8.1b.

Case 3: $n = 4$

If the sample size is increased to 4, then the sample means and probabilities are as shown in Tables 8.8 and 8.9. Comparing Table 8.9 with Tables 8.5 and 8.7 reveals that the range of the possible values of the sample mean has been further reduced, from (5–2) to (4.5–2.5) – that is, from 3 to 2. The sampling distribution shown in Fig. 8.3 looks almost like a normal distribution.

In Example 8.1, we saw how the sampling distribution can be identified, how the sample size can affect the variation of a sample mean distribution, and how the sample distribution approaches a bell-shaped normal distribution when sample size increases. It remains to consider how the sample size affects the number of possible sample means and sample variances. Let N be the size of a population with mean μ_x and standard deviation σ_x. A random sample of n observations is drawn from this population, so there are $\binom{N}{n}$ sample means $\overline{X}_i$ and $\binom{N}{n}$ sample variances S_i^2, where $i = 1, 2, \ldots, \binom{N}{n}$. Let's use the information related to Example 8.1 to illustrate this concept.

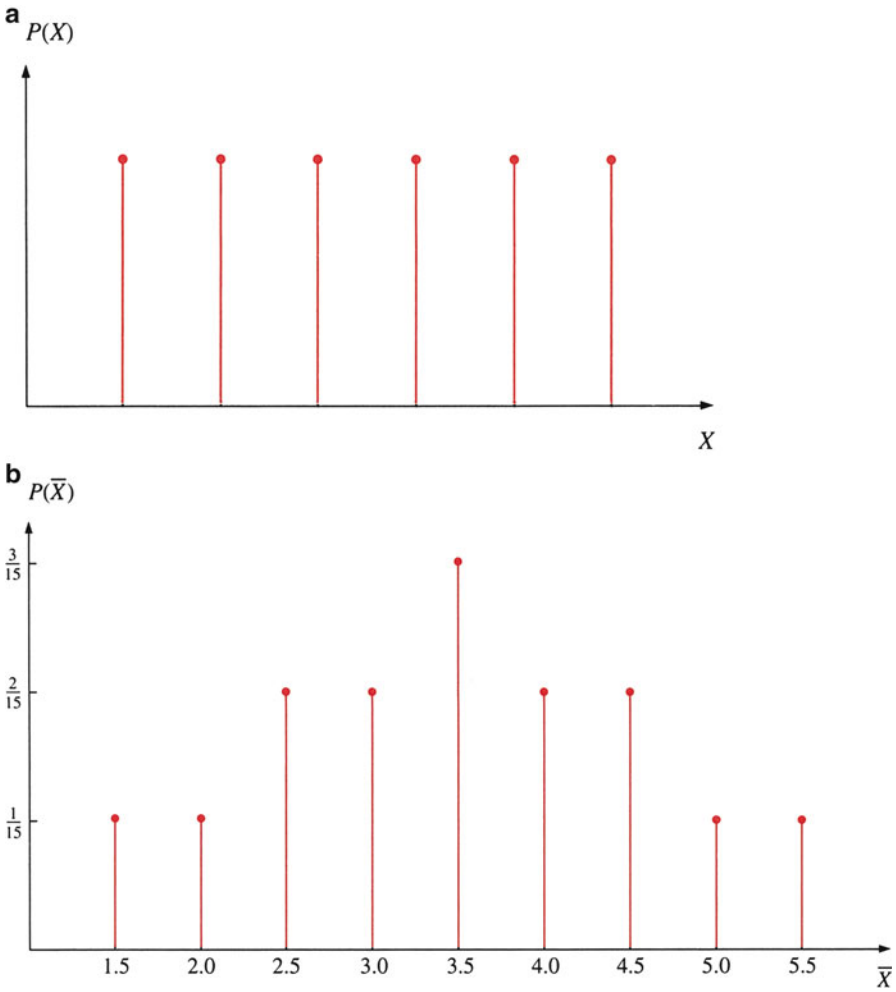


Fig. 8.1 (**a**) Population distribution. (**b**) Sampling distribution of mean work experience ($n = 2$)

Table 8.6 Possible samples and sample means ($n = 3$)

Sample	Sample mean	Sample	Sample mean
1, 2, 3	2	2, 3, 4	3
1, 2, 4	2.33	2, 3, 5	3.33
1, 2, 5	2.67	2, 3, 6	3.67
1, 2, 6	3	2, 4, 5	3.67
1, 3, 4	2.67	2, 4, 6	4
1, 3, 5	3	2, 5, 6	4.33
1, 3, 6	3.33	3, 4, 5	4
1, 4, 5	3.33	3,4, 6	4.33
1, 4, 6	3.67	3, 5, 6	4.67
1, 5, 6	4	4, 5, 6	5

Table 8.7 Probability distribution of $\overline{X}$ for $n = 3$

$\overline{X}$	$P(\overline{X})$
2	1/20
2.33	1/20
2.67	2/20
3	3/20
3.33	3/20
3.67	3/20
4	3/20
4.33	2/20
4.67	1/20
5	*1/20*
	1.00

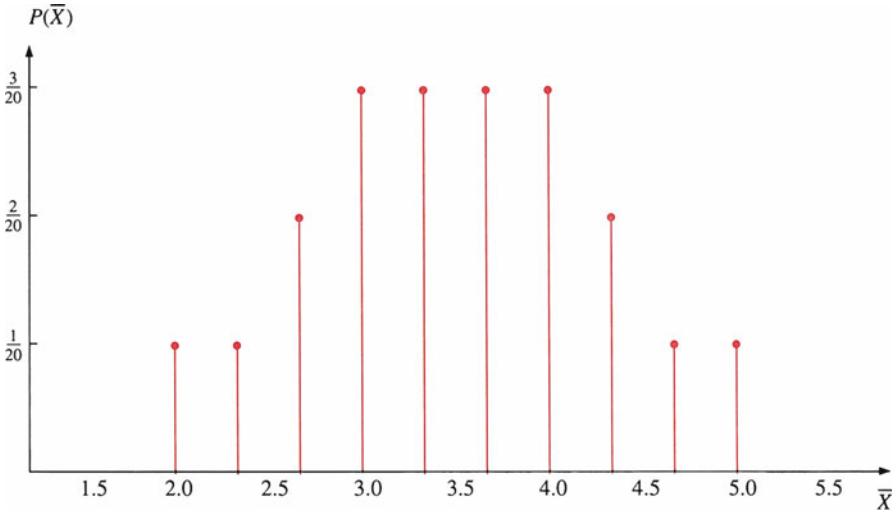


Fig. 8.2 Sample distribution of mean work experience for $n = 3$

Table 8.8 Possible samples and sample means ($n = 4$)

Sample	Sample mean	Sample	Sample mean
1, 2, 3, 4	2.5	1, 3, 5, 6	3.75
1, 2, 3, 5	2.75	1, 4, 5, 6	4
1, 2, 3, 6	3	2, 3, 4, 5	3.5
1, 2, 4, 5	3	2, 3, 4, 6	3.75
1, 2, 4, 6	3.25	2, 3, 5, 6	4
1, 2, 5, 6	3.5	2, 4, 5, 6	4.25
1, 3, 4, 5	3.25	3, 4, 5, 6	4.5
1, 3, 4, 6	3.5		

Table 8.9 Probability distribution of $\overline{X}$ for $n = 4$

$\overline{X}$	$P(\overline{X})$
2.5	1/15
2.75	1/15
3	2/15
3.25	2/15
3.5	3/15
3.75	2/15
4	2/15
4.25	1/15
4.5	*1/15*
	1.00

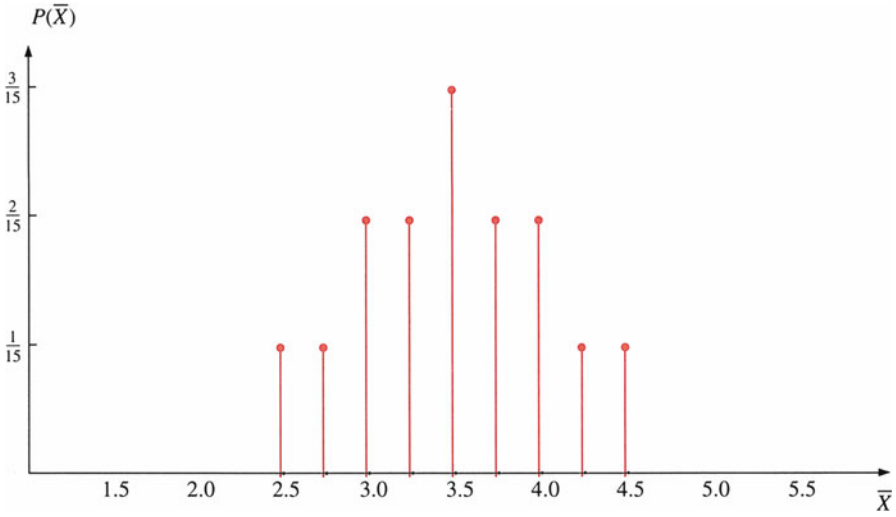


Fig. 8.3 Probability distribution $\overline{X}$ for $n = 4$

Example 8.2 Sizes of Sample Means and Their Distributions. In Example 8.1, $N = 6$, and random samples of size 2, 3, and 4 were used to show how all possible sample means can be calculated when sampling without replacement. The possible numbers of sample means and sample variances for these three alternative samples are

$$\binom{6}{2} = \frac{6!}{2!(6-2)!} = \frac{(6)(5)(4)(3)(2)(1)}{(2)(1)(4)(3)(2)(1)} = 15$$

$$\binom{6}{3} = \frac{6!}{3!(6-3)} = \frac{(6)(5)(4)(3)(2)(1)}{(3)(2)(1)(3)(2)(1)} = 20$$

$$\binom{6}{4} = \frac{6!}{4!(6-4)!} = \frac{(6)(5)(4)(3)(2)(1)}{(2)(1)(4)(3)(2)(1)} = 15$$

If sampling with replacement, the number of samples will be $(6)^2 = 36$, $(6)^3 = 216$, and $(6)^4 = 1296$.

In the next section, we will discuss the concepts of mean and variance analytically for the sample mean distribution in accordance with the results we got in Examples 8.1 and 8.2.

8.4.2 Mean and Variance for a Sample Mean

Example 8.1 shows how a random sample of n observations is drawn from a population with mean μ and variance σ_X^2, where the sample members are denoted $X_1, X_2, \ldots, X_n$. The sample mean is obtained from a random sample drawn from the population, so the expected value of the sample mean $\overline{X}$ of Eq. 8.1 is the population mean μ:[1]

$$\mu_{\overline{X}} = E(\overline{X}) = \mu \tag{8.2}$$

The variance of the sample mean is equal to the variance of the summation of the individual observations of X divided by the number of observations in the sample. This can be written arid simplified as follows:[2]

[1] This is because

$$\begin{aligned} E(\overline{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}[E(X_1) + E(X_2) + \cdots + E(X_n)] \\ &= \frac{1}{n}(n\mu) = \mu \end{aligned}$$

[2] $X_1, X_2, \ldots, X_n$ are independent of each other, so we can use Eq. 6.31 in Chap. 6 to obtain

$$\begin{aligned} \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \\ &= n\sigma_X^2 \end{aligned}$$

Therefore,

$$\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left(n\sigma_x^2\right) = \frac{\sigma_X^2}{n}$$

Because σ_X^2 generally is not known, it can be estimated by s_X^2, the sample variance:

$$s_X^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{\sigma_X^2}{n} \tag{8.3}$$

The variance of the sampling distribution of $\overline{X}$ decreases as the sample size n increases. In other words, the more observations in the sample, the more concentrated is the sampling distribution of the sample mean around the population mean, as we saw in Example 8.1. Using Eq. 8.3, we find the standard deviation of the sample mean as follows:

$$\sigma_{\overline{X}} = \sigma_X / \sqrt{n} \tag{8.4}$$

Equation 8.2 is applicable to both an infinite sample or a finite sample, with and without replacement. Equation 8.4, however, is applicable only to either an infinite sample or a finite sample with replacement.

8.4.3 Sample Without Replacement from a Finite Sample

In the case of a sample drawn without replacement, it is important to consider the size of the sample relative to the population size N. If the sample size is less than 5 % of the population ($n \leq .05\ N$), then Eq. 8.4 may be used as it appears here. If the population is large, and if the sample size is larger than 5 % of the total population ($n > .05\ N$), then a correction factor must be incorporated into Eq. 8.4. When samples are drawn from populations without replacement, each observation can be chosen only once. Therefore, as the available choices for new sample members become large, the chance that a given sample member will be chosen is still random, but there is a larger probability of its being chosen than in sampling *with* replacement because fewer members remain in the population. This has been shown to bias sample variance and standard deviation. The bias can be corrected as follows:

$$\mathrm{Var}(\overline{X}) = \frac{\sigma_X^2}{n} \cdot \frac{N-n}{N-1} \tag{8.5}$$

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \tag{8.6}$$

for all samples where $n > .05\ N$. Here, $(N - n)/(N - 1)$ is called the *finite population multiplier*.[3] Equations 8.5 and 8.6 are the variance and standard deviation in cases of finite population.

[3] We encountered this issue in Chap. 6, where we found that the hypergeometric distribution considered the population size N but the binomial distribution did not. Equation 6.15 can be redefined as

$$\begin{bmatrix}\text{Variance of hypergeometric}\\ \text{random variable}\end{bmatrix} = \begin{bmatrix}\text{Variance of corresponding}\\ \text{binomial random variable}\end{bmatrix} \cdot \left[\frac{N-n}{N-1}\right]$$

Table 8.10 All possible sample means and associated probabilities

Number of sample	Combinations of grade points	Mean $(\overline{X})$
1	1.5, 2	1.75
2	1.5, 3	2.25
3	1.5, 3.5	2.50
4	1.5, 4	2.75
5	1.5, 5	3.25
6	2, 3	2.50
7	2, 3.5	2.75
8	2, 4	3.00
9	2, 5	3.50
10	3, 3.5	3.25
11	3, 4	3.50
12	3, 5	4.00
13	3.5, 4	3.75
14	3.5, 5	4.25
15	4, 5	4.50
		$E(\overline{X}) = 3.167$

Example 8.3 Sample Mean Distribution with Samples of Different Size. Suppose a class has six students with the following grade points: 1.5, 2, 3, 3.5, 4, 5. The population mean and standard deviation of this set of data are

$$\mu = (1.5 + 2 + 3 + 3.5 + 4 + 5)/6 = 3.167$$

$$\sigma_X = \left[(1.5 - 3.167)^2 + \cdots + (5 - 3.167)^2/6\right]^{1/2} = (8.334/6)^{1/2} = 1.179$$

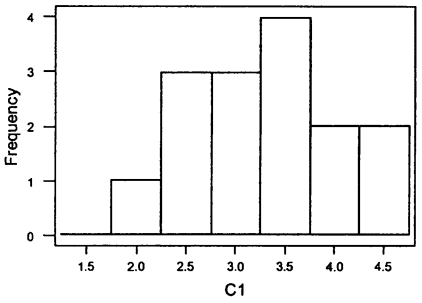
If samples of two were drawn from this population, 15 combinations would be possible. Fifteen samples of two students each and the calculated $\overline{X}$ for each sample are listed in Table 8.10. These sample means are not all 3.167, but they are close to 3.167.

The mean and standard deviation and other related information for 15 sample means generated by MINITAB are presented in Fig. 8.4. Here, we learn that (1) the average mean of these 15 values is 3.167 (this is equal to $\mu = 3.167$) and (2) the standard deviation of these 15 values is .772 or .745, which depends on whether $n - 1$ or n is used as the denominator for calculating the standard deviation. Substituting $N = 6$, $n = 2$, and $\sigma_X = 1.179$ into Eq. 8.6, we obtain

$$\sigma_{\overline{X}} = \frac{1.179}{\sqrt{2}} \sqrt{\frac{6-2}{6-1}} = .7456$$

This result also proves that Eq. 8.6 holds approximately true. We have proved the $E(\overline{X}) = \mu$, indicated in Eq. 8.2, holds true.

```
SET INTO C1
DATA> 1.75 2.25 2.50 2.75 3.25 2.50 2.75 3.00 3.50 3.25 3.50 4.00 3.75 4.25 4.50
DATA> END
MTB > GPRO
* NOTE  * Professional Graphics are enabled.
         Standard Graphics are disabled.
         Use the GSTD command to enable standard Graphics.
MTB > HISTOGRAM C1;
SUBC> MIDPOINT;
SUBC> BAR.
```




```
MTB > DESCRIBE C1
```

Descriptive Statistics

Variable	N	Mean	Median	Tr Mean	StDev	sE Mean
C1	15	3.167	3.250	3.173	0.772	0.199

Variable	Min	Max	Q1	Q3
C1	1.750	4.500	2.500	3.750

Fig. 8.4 Sample distribution on sample mean ($N = 6$, $n = 2$)

Now we offer two other examples to show how Eqs. 8.2, 8.4, and 8.6 can be applied.

Suppose an extremely large population has mean $\mu = 90.0$ and standard deviation $\sigma_X = 15.0$. We already know that the expected value of the sample mean is equal to the population mean. Therefore, the sampling distribution of the sample means for a sample size of $n = 25$ has the following parameters:

$$E(\overline{X}) = \mu = 90.0$$
$$\sigma_{\overline{X}} = \sigma_X / \sqrt{n} = 15.0 / \sqrt{25} = 15.0/5.0 = 3.0$$

Suppose this time that the population is $N = 50$ firms in an industry. Further, let's assume that the population represents earnings per share (EPS) observations for all firms in a given industry with mean \$10 and standard deviation \$2. A financial analyst takes a random sample of 20 of these firms. Because the sample size $n > .05\ N$, our estimate of the standard deviation of the sample mean must take

the correction factor $(N - n/N - 1)$ into account. We use Eqs. 8.2 and 8.6 to calculate the sample mean and sample standard deviation:

$$\begin{aligned} E(\overline{X}) &= \mu = \$10 \\ \sigma_{\overline{X}} &= \left(\sigma_X/\sqrt{n}\right)\sqrt{(N-n)/(N-1)} \\ &= \left(2\Big/\sqrt{20}\right)\sqrt{(50-20)/(50-1)} \\ &= \$0.35 \end{aligned}$$

If the population is either normally distributed or large and $n \geq 30$, then the random variable Z is distributed standard normally and is defined as follows:

$$Z = \frac{(\overline{X} - \mu)}{\sigma_X/\sqrt{n}} \tag{8.7}$$

Researchers then can use Eq. 8.7 and the standard normal distribution table (Table A3 in Appendix A at the end of the book) to do statistical analysis in terms of $\overline{X}$ and $\sigma_X/\sqrt{n}$.

Application 8.3 Probability and Sampling Distributions of Radial Tires' Lives. Suppose there is a population of radial tires whose lives are normally distributed and have a mean of 26,000 miles with a standard deviation of 3,000 miles. A random sample of 36 of these tires was taken and found to have a mean life of 25,000 miles. If the population parameters are correct, what is the probability of finding a sample mean less than or equal to 25,000? Following Sect. 7.4 of Chap. 7 on the use of the normal area table and using Eq. 8.7, we find that the probability is

$$P(\overline{X} \leq 25,000) = P\left[(\overline{X} - \mu)/\sigma_{\overline{X}} \leq P(25,000 - \mu)/\sigma_{\overline{X}}\right]$$

But, from Eq. 8.4, we know that the standard deviation of the sample mean is

$$\sigma_{\overline{X}} = \sigma_X/\sqrt{n} = 3,000\Big/\sqrt{36} = 500$$

and

$$\begin{aligned} P(\overline{X} \leq 25,000) &= P[Z \leq (25,000 - 26,000)/500] \\ &= P[Z \leq -2] \end{aligned}$$

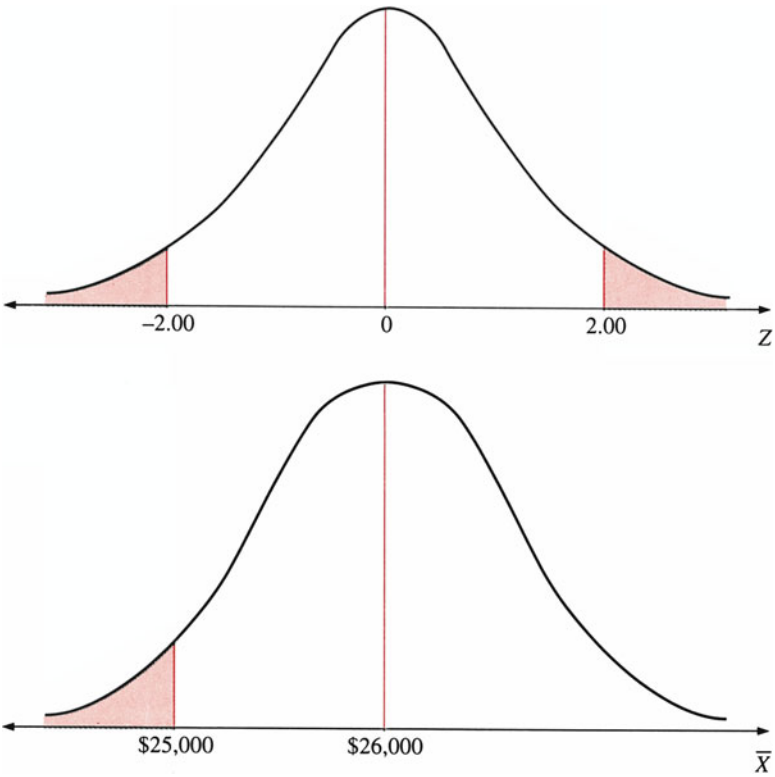


Fig. 8.5 The normal probability curve for Z and $\overline{X}$ statistics

Because the distribution of Z is a standard normal distribution, we use Table A3 to calculate the probability:

$$\begin{aligned} P\left(\overline{X} \leq 25,000\right) &= F_Z(-2.0) \\ &= 1 - F_Z(2.0) \\ &= 1 - .9772 \\ &= .0228 \end{aligned}$$

Thus, the probability that the sample mean for the life of the radial tires is less than or equal to 25,000 miles is approximately 2.3 %. The normal probability curves for the Z and $\overline{X}$ statistics for the population distribution are shown in Fig. 8.5.

To further investigate the relationship between sample size and the sampling distribution of $\overline{X}$, we use the information of Application 8.3. The expected value and standard deviation of the sampling distribution of $\overline{X}$ for five different sample sizes can be calculated as follows:

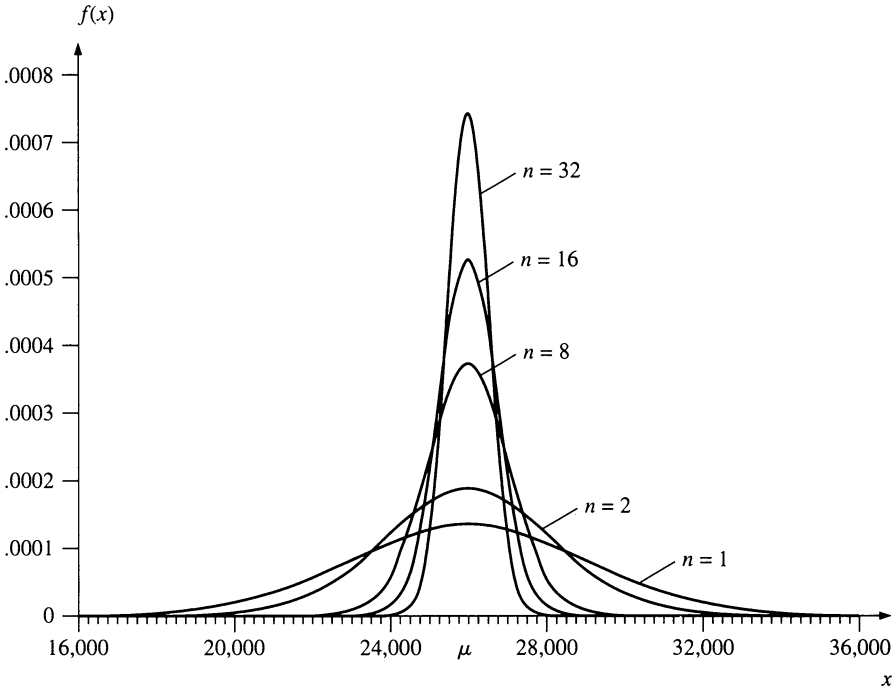


Fig. 8.6 Sampling distributions of $\overline{X}$ for five different sample sizes

Sample size	$E(\overline{X}) = \mu$	$\sigma_{\overline{X}} = \sigma_X/\sqrt{n}$
$n = 1$	26,000	$\frac{3,000}{1} = 3,000$
$n = 2$	26,000	$\frac{3,000}{\sqrt{2}} = 2121.3407$
$n = 8$	26,000	$\frac{3,000}{\sqrt{8}} = 1060.6703$
$n = 16$	26,000	$\frac{3,000}{\sqrt{16}} = 750$
$n = 32$	26,000	$\frac{3,000}{\sqrt{32}} = 530.3258$

On the basis of this information, five different normal distributions with mean 26,000 and five different standard deviations are displayed in Fig. 8.6. Sample size does not affect the expected value μ of the sample mean, but the standard deviation $\sigma_{\overline{X}}$ of the sample mean becomes smaller when the sample size increases.

8.5 Sampling Distribution of the Sample Proportion

Sometimes in statistical analysis, it is important to estimate the proportion of a certain characteristic in a population. For example, it may be of interest to estimate the proportion of people in New York City who are unemployed or the proportion of students at Rutgers University who favor changing the grading system. The *sample proportion*, $\hat{p}$, is simply the number of sample members X with the specified characteristic divided by the sample size n:

$$\hat{p} = \frac{X}{n} \tag{8.8}$$

The mean and variance of a sample proportion can be derived from the binomial distribution discussed in Chap. 6. Recall that the mean and standard deviation of a binomially distributed random variable X are

$$\mu = np \tag{6.11}$$

$$\sigma = \sqrt{np(1-p)} \tag{6.12}$$

where p is the probability of success.

From Eqs. 6.11 and 6.12, the mean and standard deviation of a sample proportion $\hat{p}$ can be calculated as

$$\mu_{\hat{p}} = \frac{np}{n} = p \tag{8.9}$$

$$\sigma_{\hat{p}} = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}} \tag{8.10}$$

For the same reasons as stated in Sect. 8.4, we need the finite population correction if $n > .05\ N$. The corrected variance for large samples (relative to population size) is

$$\sigma_{\hat{p}}^2 = [p(1-p)/n](N-n)/(N-1) \tag{8.11}$$

and

$$\sigma_{\hat{p}} = \sqrt{[p(1-p)/n}\cdot\sqrt{[(N-n)/(N-1)]} \tag{8.12}$$

Finally, if the sample size is large – say, greater than 30 – then the following Z statistic is approximately distributed as standard normal:

$$Z = (\hat{p} - p)/\sigma_{\hat{p}} \tag{8.13}$$

Mean, variance, and standard deviation calculations are performed in the same manner as in Eq. 8.7 of Sect. 8.4. The following example illustrates the inferential use of the Z statistic shown in Eq. 8.13.

Example 8.4 Calculating the Probability of Defective Chips. A shipment of 1,000 calculator chips arrives at Kraft Electronics. Suppose the company takes a random sample of 50 chips. The company claims that the proportion of defective chips in this shipment is about 25 %. Assume the claim is correct. What is the probability that this shipment will contain between 23 % and 27 % defective chips?

We begin with the information available and calculate the associated probability with Eq. 8.13. The population proportion is $p = .25$, and the related probability can be defined as

$$P(.23 < \hat{p} < .27) = P\left(\frac{.23 - p}{\sigma_{\hat{p}}} < \frac{\hat{p} - p}{\sigma_{\hat{p}}} < \frac{.27 - p}{\sigma_{\hat{p}}}\right)$$

and

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n} = \sqrt{(.25)(.75)/50} = .061$$

Therefore,

$$P(.23 < \hat{p} < .27) = P\left(\frac{.23 - .25}{.061} < \frac{\hat{p} - p}{\sigma_{\hat{p}}} < \frac{.27 - .25}{.061}\right)$$

Using the cumulative distribution function $F_Z(Z)$ of the standard normal random variable and the standard normal distribution table (Table A3 in Appendix A), we obtain

$$\begin{aligned} P(.23 < \hat{p} < .27) &= P\{-.33 < Z < .33\} \\ &= F_Z(.33) - F_Z(-.33) \\ &= F_Z(.33) - [1 - F_Z(.33)] \\ &= .6293 - [1 - .6293] \\ &= .2586 \end{aligned}$$

There is a 25.86 % chance that between 23 % and 27 % of the chips in this shipment will be defective.

8.6 The Central Limit Theorem

As we found in Sect. 8.4, the sample means of $\binom{N}{n}$ possible samples have the following properties:

1. If the population is normally distributed, the distribution of the sample mean is normal.
2. If the population is large but not normally distributed – for example, if the distribution is uniform or U-shaped – the distribution of sample mean approaches a normal distribution provided that the sample is large, as indicated in Fig. 8.7.[4]

Following these results is one of the most important theorems in statistics, the *central limit theorem:*

As the sample size (n) from a given population gets "large enough," the sampling distribution of the mean, $\overline{X}$, can be approximated by a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, regardless of the distribution of the individual values in the population.

Alternatively, the central limit theorem can be stated in the following way. Let, $X_1, X_2, \ldots X_n$ be independent and identically distributed random variables with mean μ and standard deviation σ. Let $\overline{X}$ represent the sample mean with sample size n. Then, as n becomes large, the distribution of the following Z statistic as indicated in Eq. 8.7 approaches the standard normal distribution:

$$Z = \frac{\overline{X} - \mu}{\sigma_X/\sqrt{n}} \tag{8.7}$$

Many useful calculations can be made via the central limit theorem. It is worthwhile to know that the central limit theorem can be employed to justify using the normal distribution as an approximation for both binomial and Poisson distributions, as discussed in Chap. 6.

Why is the central limit theorem so important in statistics? It enables us to analyze the means of many different random variables even when we don't know the actual population distributions of these variables. For instance, in Application 8.3, we computed the probability that the mean tire life was less than or equal to 25,000 miles. Even though we assumed that tire life was normally distributed, we could have conducted this analysis without making that assumption simply by using the central limit theorem.

Other possible uses of the central limit theorem include quality control analysis (such as examining the mean number of defective parts in a car, which will be discussed in Chap. 10), investment analysis (such as examining the mean rates of return for stocks, which was discussed in Chaps. 3 and 4), and educational analysis (such as examining mean IQ scores).

[4] Random samples from a uniform distribution for sample size $n = 2$, 5, 10, 25, and 50 are presented in Appendix 1.

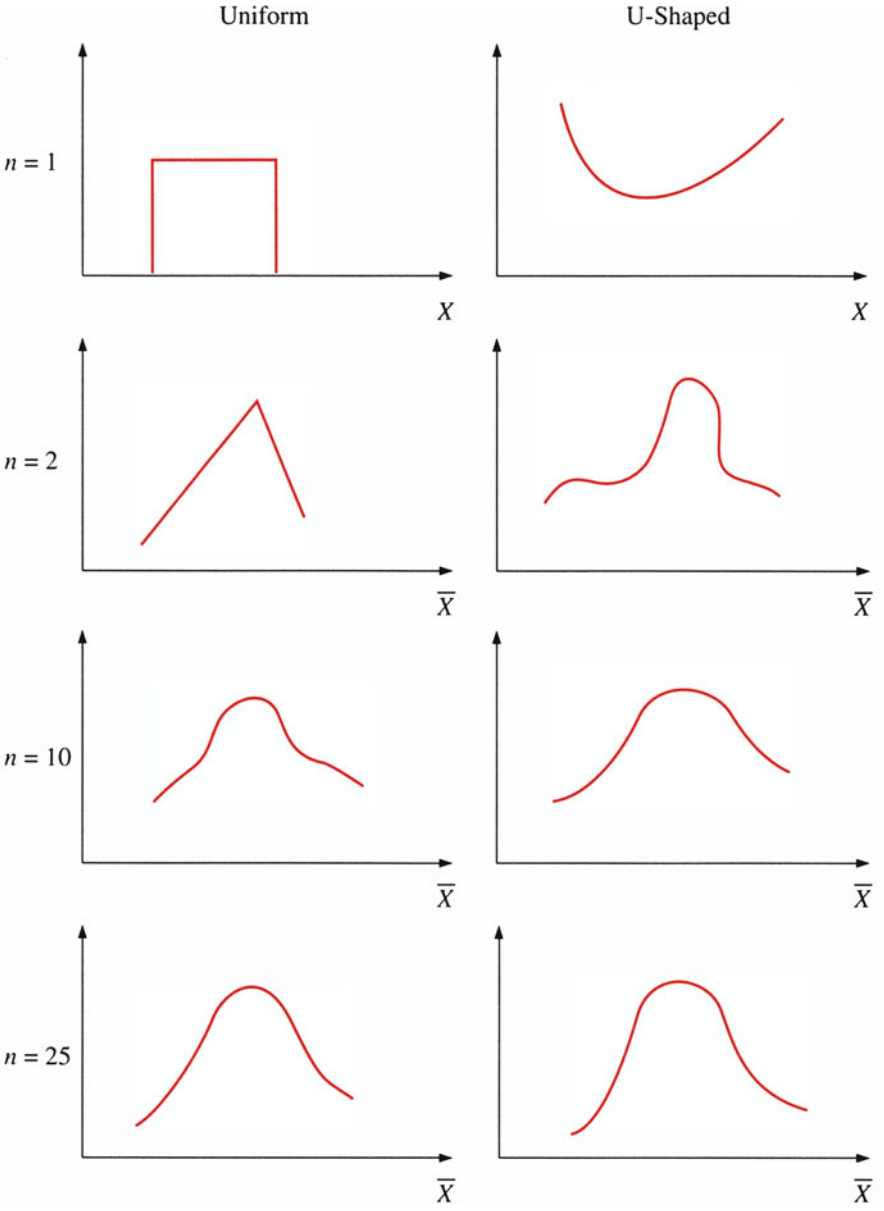


Fig. 8.7 Sample mean distribution with samples of different size

Example 8.5 Illustrating the Central Limit Theorem. The distribution of annual earnings of all marketing assistant professors in the United States with 5 years of experience is skewed negatively, as shown in part (a) of Fig. 8.8. This distribution has a mean of $55,000 and a standard deviation of $4,000. Say we draw a random sample of 50 assistant professors of marketing. What is the probability that their

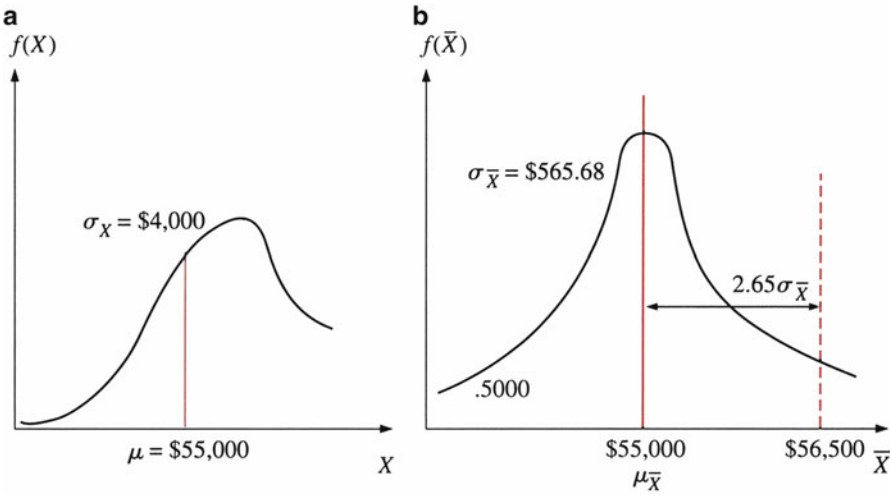


Fig. 8.8 (**a**) Population and (**b**) sampling distributions for marketing assistant professors' annual earnings

annual earnings will average more than $56,500? Part (b) of Fig. 8.8 shows the sampling distribution of the mean that will result. It also indicates the area representing "earnings over $56,500."

First we calculate the standard deviation of the mean from the population standard deviation in accordance with Eq. 8.4:

$$\begin{aligned}\sigma_{\overline{X}} &= \frac{\sigma_X}{\sqrt{n}} \\ &= \frac{\$4,000}{\sqrt{50}} \\ &= \$565.68\end{aligned}$$

From Eq. 8.2, we know that

$$E\left(\overline{X}\right) = \mu_{\overline{X}} = \mu = \$55,000$$

Because the sample mean $\overline{X}$ is normally distributed, we can use Eq. 8.7 to calculate

$$\begin{aligned}Z &= \frac{\overline{X} - \mu}{\sigma_X / \sqrt{n}} \\ &= \frac{\$56,000 - \$55,000}{\$565.68} \\ &= 2.65\end{aligned}$$

Finally, we use the Z statistics given in Table A3 in Appendix A to obtain the desired probability:

$$\begin{aligned} P(\overline{X} > \$56,500) &= P(Z > 2.65) \\ &= .5000 - .4960 = .0040 \end{aligned}$$

We have determined that there is .4 % chance of average annual earnings being more than $56,500 in a group of 50 assistant marketing professors.

8.7 Other Business Applications

Application 8.4 Audit Sampling. It is possible in accounting to make inferences about an entire large, finite population by drawing samples of size n and thus using only a small portion of the data. The information in Table 8.11 on a sample of 30 accounts was taken from the population of 3,000 trade accounts receivable for a given company.[5] Using Eq. 8.1, we can calculate the mean of the sample in Table 8.9 as follows:

$$\begin{aligned} \overline{X} &= \left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i \\ &= \left(\frac{1}{30}\right)[195.81 + 152.65 + \cdots + 215.95] \\ &= \$202.10 \end{aligned} \tag{8.1}$$

Using Eq. 4.7, we can calculate the variance of the sample as follows:

$$\begin{aligned} s^2 &= \sum_{i=1}^{\mathrm{n}} \frac{(X_i - \overline{X})^2}{\mathrm{n} - 1} \\ &= (1/29)\left[(195.81 - 202.10)^2 + (152.65 - 202.10)^2 + \cdots + (215.95 - 202.10)^2\right] \\ &= 719.164 \end{aligned}$$

Armed with this information and with information on the population standard deviation σ_X, we can make inferences about the population mean.[6] This is achieved by using the same structure as in Table 7.2.

[5] Bailey A.D. Jr.: Statistical Auditing: Review, Concepts and Problems, pp. 138–42. New York, Harcourt, Brace Jovanovich (1981)

[6] If population standard deviation is not available, we can substitute s_X for σ_X, but in this case, the Z statistics defined in Eq. 8.7 can no longer be used. A different statistic can be used, however. See Sect. 9.3 for the discussion and application.

Table 8.11 Sample of trade accounts receivable balances

Account number	Customer name	Book amount[a] $\bar{X}_1$	Rank by dollar size
101	Beekmans, F.M.	$ 195.81	10
102	Morsby, A.F.	152.65	2
103	Sack, I.E.	225.74	25
104	Hoschke, K.R.	190.73	8
105	Hosken, A.J.	207.66	18
106	Manitzky, A.A.	207.57	17
107	Worner, C.J.	210.21	19
108	Walsh, A.	147.75	1
109	Ryland, K.L.	217.73	22
110	Nolde, J.P.	206.47	15
111	Rehn, L.M.	222.12	24
112	Argent, A.	204.26	14
113	Mollison, A.M.	247.35	30
114	Conolly, E.W.J.	230.24	27
115	England, A.G.	198.12	11
116	Brown, C.	220.03	23
117	Luther, E.	216.36	21
118	Sarikas, A.D.	241.62	29
119	Martinez, B.P.	169.53	6
120	Beech, D.F.	228.98	26
121	Bedford, B.A.	159.57	4
122	Apps, A.J.	194.75	9
123	Hamlyn-Harris, T.H.	181.01	7
124	Mangan, M.R.	157.60	3
125	Topel, Z.H.	198.15	12
126	Westaway, W.R.	203.73	13
127	A-Izzedin, T.B.	206.47	16
128	Alrey, R.C.	239.12	28
129	Biment, W.	165.76	5
130	Dimick, M.C.	215.95	20
		$6,063.04	

Source: Andrew D. Bailey, Jr., *Statistical Auditing: Review, Concepts and Problems,* pp. 138–42. Copyright © 1981 by Harcourt Brace Jovanovich, Inc., reprinted by permission of the publisher
[a]These amounts were generated by using a mean of $200.00, a standard deviation of $30.00, and an assumed normal distribution. Rounding is to the nearest cent

The fact that the population mean is unknown is the motivation for this analysis. By taking a sample of 30 observations from a population of 3,000, the auditor can calculate the mean and standard deviation of this sample. From this type of information, the auditor can use the central limit theorem to determine the possible ranges that should include the true population mean if the population standard deviation is known. Suppose the population standard deviation is $25,560. We can use Eq. 8.4 to estimate the standard deviation of the sample mean:[7]

[7] If the population standard deviation is not known, then we can use the information on sample mean and sample variance to do a similar analysis. This kind of analysis will be done in Sect. 9.3.

Table 8.12 Confidence intervals of accounts receivable population mean (μ)

Confidence level (%)	Confidence interval
68.0	$\$197.43 < \mu < \$206.77$
95.5	$\$192.77 < \mu < \$211.43$
99.7	$\$188.10 < \mu < \$216.10$

$$S_{\overline{X}} = \sigma_X / \sqrt{n} = 25.56 / \sqrt{30} = 4.667$$

The guidelines listed in Table 7.2 give us a rule for determining *confidence intervals*. We can use this rule to make the three confidence-interval statements listed in Table 8.12. For example, the first confidence-interval statement can be expressed as follows: "We can be about 68 % confident that the population mean μ will fall between \$197.43 and \$206.77."

Application 8.5 Patient Waiting Time. Sloan and Lorant (1977) studied the relationship between the length of time patients wait in a physician's office and certain demand and cost factors.[8] They obtained data on typical patient waiting times for 4,500 physicians and reported a mean waiting time of 24.7 min and a standard deviation of 19.3 min.

Suppose a pediatrician does not have this set of data and has one of the nurses in the office monitor the waiting times for 64 randomly selected patients during the year. Applying the central limit theorem, we know that the sample mean, $\overline{X}$, is approximately normally distributed and that the mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$ are

$$\mu_{\overline{X}} = \mu = 24.7 \text{ min}$$
$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{19.3}{\sqrt{64}} = 2.4 \text{ min}$$

The chance of the sample mean falling between 18 and 26 min can be calculated as follows. Because $\overline{X}$ is normally distributed, we can use Eq. 8.7 to calculate

$$Z_1 = \frac{\overline{X} - \mu}{\sigma_X / \sqrt{n}} = \frac{18 - 24.7}{2.4} = -2.79$$
$$Z_2 = \frac{26 - 24.7}{2.4} = .54$$

By using Table A3 in Appendix A, we can calculate the probability that the sample mean $\overline{X}$ falls between 18 and 26 min as

[8] Sloan F.A., Lorant J.H.:The role of patient waiting time: Evidence from physicians' practices. J. Bus. , October, 486–507 (1977)

$$\begin{aligned} P\left(18 \leq \overline{X} \leq 26\right) &= P(-2.79 \leq Z \leq .54) \\ &= .4974 + .2054 \\ &= .7028 \end{aligned}$$

There is about a 70.28 % chance that the sample mean will fall between 18 and 26 min. The pediatrician can use this information to determine how efficiently the office is operating.

8.8 Summary

In this chapter, we began our treatment of inferential statistics by discussing the concept of sampling and sampling distributions. Inferential statistics deals with drawing inferences about population parameters by looking at a sample of the population. We considered the costs and benefits of sampling and how to draw a random sample. In addition, we discussed the distribution of the sample mean and one of the most important theorems in statistics, the central limit theorem.

In Chap. 9, we will examine other important continuous distributions. In Chap. 10, we continue our discussion of inferential statistics by introducing the concepts of point estimation and confidence intervals.

Questions and Problems

1. A grocery store sells an average of 478 loaves of bread each week. Sales (X) are normally distributed with a standard deviation of 17.
   (a) If a random sample of size $n = 1$ (week) is drawn, what is the probability that the $\overline{X}$ value will exceed 495?
   (b) If a random sample of size $n = 4$ (weeks) is drawn, what is the probability that the $\overline{X}$ value will exceed 495?
   (c) Why does your response in part (a) differ from that in part (b)?
2. A random variable s measures the daily balances in customers' savings accounts. It is normally distributed, with a mean of μ_s – \$108 and a standard deviation of \$15.
   (a) If a random sample of size $n = 4$ is drawn, what is the probability that the s_4 value exceeds \$116?
   (b) If a random sample of size $n = 16$ is drawn, what is the probability that the s_{16} value exceeds \$116?
   (c) What happened to the standard deviation of $\bar{s}$ when the sample size increased from $n = 4$ to $n = 16$?
   (d) What happened to the probability of observing $\bar{s} \geq \$116$ as the sample size increased from $n = 4$ to $n = 16$?

3. On average, a book distributor fills orders for 1,000 books per day. If daily orders are normally distributed and the standard deviation is 100, what is the probability that a 5-day average will be between 900 and 1,100 books?
4. A company makes a pastry called a chocco. During the manufacturing process, the individual choccos are placed in a baking oven. The time it takes to bake them is normally distributed around a mean of 64 min with a standard deviation of 5 min. Thus, distribution of the population of baking times is normal in shape.

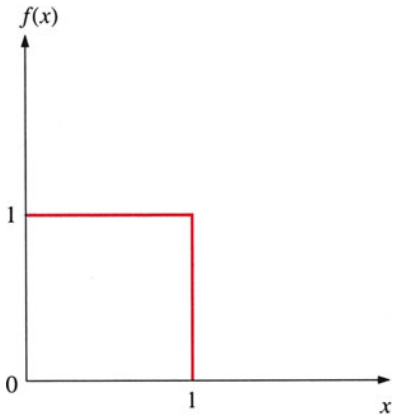
   (a) When choccos are baked, what is the probability that the mean baking time of four choccos will be 64 min and 45 s or longer?
   (b) What proportion of the individual choccos bake in 57 min or less?

5. In a large group of corporate executives, 20 % have no college education, 10 % have exactly 2 years of college, 20 % have exactly 4 years, and 50 % have 6 years. A sample size of 2 (with replacement) is to be taken from this population. Find the sampling distribution of the mean number of years in college of the executives in the sample.
6. Out of 10 pay telephones located in a municipal building, two phones are to be picked at random, with replacement, for a study of phone use. The actual usage of the phones on a particular day is shown in the accompanying table.

Number of calls	Number of phones with this number of calls
10	2
12	5
16	3

   (a) Find the sampling distribution of the average number of calls per phone in the sample of two phones.
   (b) Find the variance of this distribution.

7. To demonstrate the central limit theorem, draw 100 samples of size 5 from a random-number table and calculate the sample mean for each of the 100 samples. Construct a frequency distribution of sample means. Do the same for 100 samples of size 10 and compare the two frequency distributions. Does the central limit theorem appear to be working?
8. The accompanying probability density function is a uniform distribution showing that a certain delicate new medical device will fail between 0 and 10 years after it is implanted in the human body. The mean time to failure is $\mu = 5$ years, and the standard deviation is $\sigma = 2.88$ years.



(a) Verify that the total area beneath this density function is 1.0.
(b) Find the probability that an individual device will fail more than 8 years after implantation.
(c) Find the probability that in a sample of $n = 36$ of these devices, the sample mean time of failure $\bar{x}$ will be 8 years or less.

9. The daily catch of a small tuna-fishing fleet averages 130 t. The fleet's logbook shows that the weight of the catch varies from day to day and this variation is measured by the standard deviation of the daily catch, $\sigma = 42$ t. What is the probability that during a sample of $n = 36$ fishing days, the total weight of the catch will be 4,320 t or more?
10. A type of cathode ray tube has a mean life of 10,000 h and a variance of 3,600. If we take samples of 25 tubes each and for each sample we find the mean life, between what limits (symmetric with respect to the mean) will 50 % of the sample means be expected to lie?
11. The population of times measured by 3-min egg timers is normally distributed with $\mu = 3$ min and $\sigma = .2$ min. We test samples of 25 timers. Find the time that would be exceeded by 95 % of the sample means.
12. A light bulb manufacturer claims that 90 % of the bulbs it produces meet tough new standards imposed by the consumer protection agency. You just received a shipment containing 400 bulbs from this manufacturer. What is the probability that 375 or more of the bulbs in your shipment meet the new standards? (*Hint:* Use the continuity approximation.)
13. At the beginning of every decade, the US government conducts a census. Why does it take a census? What are the advantages of a census over a sample?
14. Briefly explain the relationship between inferential statistics and sampling.
15. Suppose a town consists of 2,000 people, 1,100 of whom are registered voters. You are interested in how the people in this town will vote on a bond issue. What group constitutes the population? Give an example of a sample from this population.

16. What is sampling error? Give an example. Is there any way to eliminate sampling error?
17. What is nonsampling error? Give an example. Is there any way to eliminate nonsampling error?
18. State whether each of the following represents sampling or nonsampling error.

    (a) The sample weight of newborn babies is taken with a scale that is inaccurate by 1 lb.
    (b) Sample data suggest that the price of a new home in New Jersey is $150,000 when the actual new home price is $ 180,000.
    (c) A movie theater owner asks the first 100 people leaving the theater whether they liked the movie. By chance, however, the first 100 people to leave are all women. (This in itself may say something about the movie!)

19. What is a representative sample? Why is getting a representative sample important?
20. Briefly explain the relationship between sampling cost and sampling error. Give some examples of sampling costs.
21. The mean life of light bulbs produced by the Brite Lite Bulb Company is 950 h with a standard deviation of 225 h. Assume that the population is normally distributed. Suppose you take a random sample of 12 light bulbs.

    (a) What is the mean of the sample mean life?
    (b) What is the standard deviation of the sample mean?

22. Suppose the mean amount of money spent by students on textbooks each semester is $175 with a standard deviation of $25. Assume that the population is normally distributed. Suppose you take a random sample of 25 students.

    (a) What is the mean of the sample mean amount spent on textbooks?
    (b) What is the standard deviation of the sample mean?

23. The Better Health Cereal Company produces Healthy Oats cereal. The true mean weight of a box of cereal is 24 oz with a standard deviation of 1 oz. Assume the population is normally distributed. Suppose you purchase eight boxes of cereal.

    (a) What is the mean of the sample mean weight?
    (b) What is the standard deviation of the sample mean?

24. Explain the relationship between a probability distribution and a sampling distribution.
25. Suppose the average time a customer waits at the check-out line in a grocery store is 12 min with a standard deviation of 3 min. If you take a random sample of five customers, what is the probability that the average check-out time will be at least 10 min? What is the mean of the sample check-out time? What is the standard deviation of the sample mean?
26. Historically, 65 % of the basketball players from Slam Dunk University graduate in 4 years. If a random sample of 50 former players is taken, what

proportion of the samples is likely to have at least 25 basketball players graduating?

27. A coffee machine is set so that it dispenses a normally distributed amount of coffee with a mean of 6 oz and a standard deviation of .4 oz. Samples of 12 cups of coffee are taken. What is the probability that the sample means will be more than 6.2 oz?
28. Suppose that historically 61 % of the companies on the NYSE have prices that go up each year. If random samples of 100 stocks are taken, what proportion of samples is likely to have between 55 % and 65 % of stock prices going up?
29. Suppose the mean life for a company's batteries is 12 h with a standard deviation of 3 h. If you take a sample of 20 batteries, what is the standard deviation of the sampling distribution of the mean?
30. The mean useful life of better traction tires is 40,000 miles with a standard deviation of 4,000 miles. If you purchase four of these tires for your car, what is the probability that the mean useful life of the four tires is less than 35,000 miles?
31. The mean interest rate of 500 money market mutual funds is 7.98 % with a standard deviation of 1.01 %. Suppose you draw a sample of 25 mutual funds.

    (a) What is the mean of the sample mean rate?
    (b) What is the variance of the sample mean?
    (c) What is the probability that this sample will have a mean rate above 8.2 %?

32. Of 500 students in a high school, 72 % have indicated that they are interested in attending college. What is the probability of selecting a random sample of 50 students wherein the sample proportion indicating interest in college is greater than 80 %?
33. The professor in a statistics course takes a random sample of 100 students from campus to determine the number in favor of multiple-choice tests. Suppose that 50 % of the entire college population are actually in favor of the multiple-choice test. What is the probability that more than 50 % of the students sampled will favor the multiple-choice test?
34. Suppose 40 % of the students in Genius High School scored above 650 on the math portion of the SAT. What is the probability that more than 50 % of a random sample of 150 students will score less than 650?
35. The Sorry Charlie Tuna Company produces canned tuna fish. The true mean weight of a can of tuna is 6 oz with a standard deviation of 1 oz. Assume the population is normally distributed, and suppose you purchase 9 cans of tuna.

    (a) What is the mean of the sample mean weight?
    (b) What is the variance of the sample mean?

36. Suppose the time a customer waits in line at a bank averages 8 min with a standard deviation of 2 min. In a random sample of five customers, what is the probability that the average time in line will be at least 10 min? What is the mean of the sample waiting time? What is the standard deviation of the sample mean?

37. In Freeport High School, 40 % of the seniors who are eligible to vote indicated that they plan to vote in the upcoming election. What is the probability of selecting a random sample of 400 students with a sample proportion of voting greater than 35 %?
38. A credit card company accepts 70 % of all applicants for credit cards. A random sample of 100 applications is taken.

    (a) What is the probability that the sample proportion of acceptance is between .60 and .80?
    (b) What is the probability that the sample proportion is greater than .75?
    (c) What is the probability that the sample proportion is less than .65?

39. From past history, a bookstore manager knows that 25 % of all customers entering the store make a purchase. Suppose 200 people enter the store.

    (a) What is the mean of the sample proportion of customers making a purchase?
    (b) What is the variance of the sample proportion?
    (c) What is standard deviation of the sample proportion?
    (d) What is the probability that the sample proportion is between .25 and .30?

40. Suppose 60 % of the members in a lifeguards' union favor certification tests for lifeguards. If a random sample of 100 lifeguards is taken, what is probability that the sample proportion in favor of certification tests is greater than 70 %?
41. A bank knows that its demand deposits are normally distributed with a mean of $1,122 and a standard deviation of $393. A random sample of 100 deposits is taken.

    (a) What is the probability that the sample mean will be greater than $1,000?
    (b) Compute the mean of the sample mean demand deposits.
    (c) Compute the variance of the sample mean.

42. A company claims that its accounts receivable follow a normal distribution with a mean of $500 and a standard deviation of $75. An auditor will certify the bank's claim only if the mean of a random sample of 50 accounts lies within $25 of the mean. Assume that the bank has accurately reported its mean accounts receivable. What is the probability that the auditor will certify the bank's claim?
43. Consider the members of a group with ages 23, 19, 25, 32, and 27. If a random sample of two is to be taken without replacement, what is the sampling distribution for their mean age? What is the mean and variance for the distribution?
44. Answer question 31 again, assuming that the sample is taken with replacement.
45. Consider a population of six numbers, 1, 2, 3, 4, 5, and 6. What is the mean of this population? Suppose you roll a pair of dice. Construct a table showing the different possible combinations of the two numbers you will obtain. Construct a probability function for this sample. Find the mean of the sample.

46. Answer question 45 again, assuming that the sample is taken from a population of four numbers, 1, 2, 3, and 4.
47. Answer question 45 again, assuming that the sample is taken from a population of three numbers, 1, 2, and 3.
48. Compute the values for $\binom{N}{n}$ if

    (a) $N = 5, n = 2$
    (b) $N = 6, n = 3$
    (c) $N = 6, n = 2$
    (d) $N = 4, n = 3$

49. Why are we interested in the sample mean and its distribution?
50. Consider the members of a weight-loss group who weigh 225, 231, 195, 184, and 131 lb. If a simple random sample of size 2 is to be taken without replacement, what is the sampling distribution for their mean weight? What are the mean and variance for the distribution?
51. Suppose there are 2,000 members in a construction workers' union and 40 % of the members favor ratifying the union contract. If a random sample of 100 construction workers is taken, what is the probability that the sample proportion in favor of ratifying the contract is greater than 50 %?
52. From past history, a service manager at Honest Abe's Auto Dealership knows that 35 % of all customers entering the dealership will have service work done that is under warranty. Suppose 200 people enter the dealership for service work on their cars.

    (a) What is the mean of the sample proportion of customers having work done that is covered by the warranty?
    (b) What is the variance of the sample proportion?
    (c) What is standard error of the sample proportion?
    (d) What is the probability that the sample proportion is between .25 and .40?

53. A quality control engineer knows from past experience that the mean weight for ball bearings is 7.4 oz with a standard deviation of 1.2 oz. Suppose the engineer draws a random sample of 20 ball bearings. What is the probability that the mean of the sample will be greater than 8.0 oz?
54. Suppose you take an ordinary deck of 52 cards randomly select five cards without replacement. How many different combinations of sample car can you have?
55. Suppose you draw three balls without replacement from a bag of balls numbered 1–10. How many different possible combinations sample balls you have?

56. Suppose the ages of members of a senior citizens' bridge club are 63, 71, 82, 60, 84, 75, 77, 65, and 70.

    (a) Compute the population mean and standard deviation for the age of the bridge club members.
    (b) If you were to select a sample of four members from the bridge club, how many possible samples could you select?

57. Use the information given in question 43 to randomly select five samples of four people and determine the mean and standard deviation for each sample
58. In each of the following cases, find the mean a standard deviation of the sampling distribution for the sample mean, for a sample of size n from a population with mean μ and standard deviation σ.

    (a) $n = 5, \mu = 10, \sigma = 2$
    (b) $n = 10, \mu = 10, \sigma = 3$
    (c) $n = 10, \mu = 5, \sigma = 3$
    (d) $n = 20, \mu = 5, \sigma = 2$

59. Suppose the cost of sampling is 50 cents per observation. If the population has zero variance, large a sample should be collected to estimate mean of the population?
60. Suppose you would like to randomly select four of the following six companies for a study: IBM, Apple Computer, AT&T, MCI, Ford, and Chrysler. What is the probability that Apple Computer will be in the sample? What is the probability that at least one company from the auto industry, one from the computer industry, and one from the telecommunications industry will be included in the sample?
61. Suppose a population is normally distributed. What is the probability that the sample mean will be less than the population mean?
62. Suppose an obstetrician knows from past experience that the mean weight of a newborn baby is 7.5 lb with a standard deviation of 2 lb. The doctor randomly chooses five newborn babies. What is the expected value of the sample mean weight? What is the expected value of the sample mean variance?
63. Review the information given in question 62. What is the probability that a sample of 50 babies will have a mean weight greater than 8 lb?
64. A cigarette manufacturer came up with a new brand of cigarettes called Long Life. The nicotine content of the cigarettes follows a normal distribution with a mean of 20 and a standard deviation of 5. A consumer bought a pack of Long Life that contains 25 cigarettes. Consider these 25 cigarettes as a random sample.

    (a) What is the probability that a cigarette contains over 23 units of nicotine?
    (b) What is the probability that the average nicotine content for the whole pack of cigarettes is higher than 23?

65. Table 8.5 shows the probability distribution of X for $n = 2$. Show that the average of the random variable X is 3.5. What is the standard deviation of X?

66. Assume the tips received by five waitresses in a given weeknight are \$25, \$27, \$28, \$29, and \$30. We draw two numbers randomly and take the average. Write the probability distribution of the sample mean. What are the expected value and standard deviation of the sample mean?
67. In a big university, 70 % of the faculty members like to give plus and minus grades (such as B plus and C minus). The other 30 % of the faculty members do not like the plus and minus system. The school newspaper randomly surveyed 200 faculty members for their opinions. What is the probability that more than half of the faculty members interviewed will be in favor of plus and minus grades? What is the expected number of faculty interviewed who answer the question positively?
68. Assume that the amount of milk in a 16-oz bottle follows a normal distribution with a mean of 16 and a standard deviation of 1. A consumer protection agency bought 30 bottles of milk and weighed them. What is the probability that the average weight of these 30 bottles of milk falls between 15.9 and 16.1 oz?
69. If, in question 68, 90 % of the bottles contain more than 16 oz of milk, what is the probability that fewer than 3 of the 30 bottles that the agency bought contain more than 16 oz of milk?
70. The newly produced 1992 Honda boasts 45 miles per gallon on the highway. Assume that the distribution of the miles per gallon is a normal distribution with a mean of 40 and a standard deviation of 5. The Environmental Protection Agency randomly draws 100 1992 Hondas to test-drive.

    (a) What is the probability that a certain car can achieve 45 miles per gallon?
    (b) What is the probability that the average of 100 cars exceeds 45 miles per gallon?

71. In question 70, what is the probability that of the 100 cars test-driven, more than 35 cars get more than 45 miles per gallon? How many of the 100 cars tested would you expect to get more than 45 miles per gallon?
72. The National Treasury Bank wants to approve, at random, two of five loan applications that have been submitted. The loan amounts are \$5,000, \$8,000, \$9,000, \$10,000, and \$12,000. Obtain the sampling distribution of average loans. What is the expected amount of loans?
73. Recently the State Education Department of New Jersey wanted to determine the competence in math of the state's fourth-grade students. Assume that 20 % of the students are actually incompetent in mathematics. A test was given to 120 fourth-grade students in New Jersey. What is the probability that at least 20 % of the students who took the test failed it?
74. Assume that 80 % of the employees are union members, whereas 20 % are not. In the last year, 100 of 500 employees were randomly selected to receive a working bonus. If the company does not discriminate against the union members, what is the probability that 30 or more bonus recipients are union members?
75. Suppose the sampling distribution of a sample mean that was developed from a sample of size 40 has a mean of 20 and a standard deviation of 10. Assuming that the population exhibits a normal distribution, find the mean and standard deviation of the population distribution.

76. A natural food company is marketing a new yogurt that it advertises as having only half the fat of regular yogurt. The average amount of fat in a cup of regular yogurt is 1 unit. The Food and Drug Administration has asked us to investigate the product to see whether the company has engaged in false advertising. The test results are as follows:

Amount of yogurt tested	400 cups
Average amount of fat contained per cup	.52 units
Number of cups containing more than half the fat of regular yogurt	12 cups
Standard deviation of the amount of fat	.2 units

What is the probability of our observing .52 units of average fat, as shown in the report if the population average fat is .5, as stated in the advertisement?

77. On the basis of your answer to question 76, do you believe the advertisement is accurate?
78. The company in question 76 further claims that only 2 % of the cups contain more than half the fat of regular yogurt. What is the probability of our seeing more than 12 cups out of 400 (which is what we saw in the report) that contain more than half the fat of regular yogurt?
79. In a game, a player rolls two dice and counts how many points he gets between them. Write out the sampling distribution.
80. What are the expected value and standard deviation of the random variable generated in question 79?
81. In a local factory, 20 % of the assembly line workers make $5 per hour and 80 % earn $8 per hour. The union computes the mean hourly wage by randomly drawing five workers. Write the sampling distribution for the five workers' average wage.
82. Write out the sampling distribution for rolling a die and flipping a coin.
83. Suppose you play a game in which you flip three coins. If the flip is a head, you receive 1 point; if the flip is a tail, you receive 2 points. Write out the sampling distribution. What are the expected value and standard deviation of this random variable?
84. Suppose you draw two cards from a standard 52-card deck with replacement. Write out the sampling distribution for the suit drawn.
85. The MINITAB output in the figure (see pages 371–372) is 20 random samples drawn from a uniform distribution between 0 and 1. Calculate the sample means and sample standard deviations by using the MINITAB program.
86. Use MINITAB to draw histograms for both the sample means and the sample standard deviations, which have been calculated in question 85. Explain the results.
87. Use the results you got in question 85 to plot sample means against sample standard deviation. What is the probability of the range, the sample means between .45 and .55, and the sample standard deviation between .2 and .35?
88. In a survey by the United Airlines of 100 flights between Jan. 2, 2012 to Feb. 15, 2012, whether the flights arrive their destination on time or not are recorded. What is the sample proportion? Please estimate the standard

deviation of the sample proportion, and calculate the probability that the sample proportion is less than 30 %.

The financial data of 37 companies in the communication and internet sector is given below. In the data set, the debt-to-asset, dividend per share, current ratio, fixed asset turnover, ROA, and P/B ratio of September, 2011 are given.

Debt-to-asset (%)	Dividend per share	Current ratio	Fixed asset turnover	ROA	P/B
60.35	0	130.4	2.16	−5.36	0.92
78.74	0	87.75	9.8	−1.4	0.91
24.3	5.73	36.76	1.08	1.24	1.03
45.53	10.08	151.46	14.43	2.55	1.05
22.58	0	171.58	1.17	0.02	0.85
13.27	5.44	174.97	0.17	2.78	2.19
53.93	5.99	137.16	3.18	0.22	1.56
28.8	12.1	150.26	13.8	1.06	1.17
53.05	0	125.28	0.91	1.45	0.72
35.8	4.4	215.22	12.08	4.5	4.82
17.58	1.07	394.77	0.29	3.71	2.47
17.69	8.16	438.73	3.08	4.52	1.47
54.66	0	220.43	1.04	−9.68	1.88
63.65	5.06	129.16	9.33	7.41	6.5
19.89	7.6	419.59	0.48	−1.31	0.77
34.34	1.01	122.9	1.06	1.23	0.65
46.59	4.37	39.16	0.47	4.04	6.24
24.21	1.5	373.75	10.75	−1.62	0.83
33.55	2.29	253.04	9.52	2.85	1.88
40.79	7.96	176.64	9.87	4.67	2.24
35.49	7.35	218.21	6.83	2.23	1.06
39.01	0	164.8	2.02	0.26	0.84
37.82	3.1	209.99	5.42	2.57	0.94
51.42	5.52	157.65	27.38	0.73	1.49
0.32	3.42	151.9	0	0.59	0.89
22.85	5.49	54.93	0.47	2.69	2.13
45.63	3.4	155.36	4.23	1.12	0.94
57.04	4.69	135.61	5.36	2.11	1.84
4.17	11.24	1,229.48	2.28	3.16	0.9
60.48	1.06	127.84	1.9	1.4	0.65
37.98	7.65	195.28	19.95	2.87	1.3
13.01	7.26	664.24	1.41	3.28	1.1
28.33	0	111.58	0.86	−1.86	0.56
49.46	4.38	166.77	4.26	3.1	2.44
35.55	6.59	246.69	1.98	5.96	2.01
27.64	1.33	264.17	11.09	−1.58	2.46
64.18	0	121.05	25.71	2.68	1.64

Please use the sample means and variances of the above six financial variables in 2011 as an estimate of the population means and variances of the year 2012, and answer the following questions:

89. What is the probability that the sample mean of dividend per share of the 37 companies is greater than eight in 2012?
90. What is the probability that the sample mean of ROA of the 37 companies is greater than −1.5 but less than 4.5 in 2012?
91. What is the sample proportion of the ROAs of the 37 companies that are negative in 2011?
92. What is the probability that the sample proportion of negative ROAs is greater than 0.2 in 2012?

MINITAB for question 85

```
MTB >  RANDOM 10 C1;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C2;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C3;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C4;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C5;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C6;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C7;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C8;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C9;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C10;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C11;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C12;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C13;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C14;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C15;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C16;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C17;
SUBC>  UNIFORM A = 0 B = 1.
MTB >  RANDOM 10 C18;
```

(continued)

(continued)

```
SUBC>                                   UNIFORM A = 0 B = 1.
MTB >                                   RANDOM 10 C19;
SUBC>                                   UNIFORM A = 0 B = 1.
MTB >                                   RANDOM 10 C20;
SUBC>                                   UNIFORM A = 0 B = 1.
```

MINITAB for Question 85 (Continued)

```
MTB > PRINT C1-C20
```

Data display

Row	C1	C2	C3	C4	C5	C6	C7
1	0.697491	0.908023	0.029563	0.658749	0.893156	0.260689	0.253072
2	0.795475	0.874655	0.132406	0.541380	0.313101	0.435039	0.231828
3	0.087242	0.109459	0.646564	0.587851	0.510025	0.679080	0.431409
4	0.434503	0.378682	0.398421	0.174592	0.906581	0.841402	0.552284
5	0.918639	0.022752	0.699634	0.694702	0.930529	0.521381	0.776129
6	0.916262	0.653844	0.040985	0.064612	0.215344	0.785744	0.922679
7	0.273922	0.842171	0.895335	0.001518	0.281558	0.499460	0.068693
8	0.170415	0.090264	0.478746	0.440146	0.082488	0.649124	0.485246
9	0.592387	0.268102	0.222450	0.258805	0.133108	0.453357	0.600180
10	0.206343	0.353241	0.845340	0.079208	0.043057	0.242360	0.289269

Row	C8	C9	C10	C11	C12	C13	C14
1	0.397762	0.289712	0.488420	0.041530	0.999002	0.006478	0.087947
2	0.714043	0.147901	0.808523	0.143075	0.443159	0.483238	0.298676
3	0.632163	0.805329	0.098366	0.859493	6.642793	0.290319	0.746221
4	0.845410	0.265249	0.495131	0.385223	0.760022	0.436757	0.899756
5	0.823078	0.371113	0.549316	0.116782	0.980880	0.280550	0.656451
6	0.162076	0.563014	0.556136	0.103806	0.611204	0.103753	0.371799
7	0.590319	0.779153	0.296261	0.465100	0.479442	0.888985	0.248135
8	0.526598	0.558167	0.035587	0.666268	0.086061	0.714802	0.107576
9	0.188762	0.566992	0.116197	0.064171	0.510456	0.775933	0.397762
10	0.286430	0.743237	0.729364	0.171575	0.510366	0.227915	0.913113

Row	C15	C16	C17	C18	C19	C20
1	0.421177	0.649618	0.436579	0.926484	0.908857	0.158414
2	0.228157	0.405771	0.933013	0.865197	0.785487	0.987785
3	0.810667	0.515916	0.578425	0.824179	0.974714	0.093386
4	0.698915	0.747788	0.163641	0.992973	0.976558	0.302980
5	0.116913	0.492599	0.228659	0.565895	0.387054	0.525864
6	0.953395	0.855334	0.676257	0.168689	0.134300	0.763466
7	0.209950	0.415589	0.644835	0.230382	0.072556	0.096998
8	0.708988	0.744951	0.921576	0.520743	0.261274	0.808271
9	0.471597	0.970680	0.521095	0.258895	0.213202	0.524599
10	0.418113	0.158896	0.594759	0.263435	0.324730	0.251105

Appendix 1: Sampling Distribution from a Uniform Population Distribution

To show how sample size can affect the shape and standard deviation of a sample distribution, consider samples of size $n = 2$, 5, 10, 25, and 50 taken from the uniform distribution shown in Fig. 8.9.

To generate different random samples with different sample sizes, we use the MINITAB random variable generator with uniform distribution. Portions of this output are shown in Fig. 8.1b in the text discussion. First we generate 40 random samples with a sample size of 2. Similarly, we generate 40 random samples for $n = 5$, $n = 10$, $n = 25$, and $n = 50$.

Forty sample means for sample sizes equal to 2, 5, 10, 25, and 50 are presented in Table 8.13. Histograms based on the five sets of data given in Table 8.13 are presented in Figs. 8.10, 8.11, 8.12, 8.13, and 8.14, respectively. The means associated with Figs. 8.10, 8.11, 8.12, 8.13, and 8.14 are .4458, .4857, .4776, .48688, and .49650, respectively; the standard deviations associated with Figs. 8.10, 8.11, 8.12, 8.13, and 8.14 are .1927, .1300, .0890, .06235, and .04414. By comparing these five figures, we can draw two important conclusions. First, when sample size increases from 2 to 50, the shape of the histogram becomes more similar to the bell-shaped normal distribution. Second, as the sample size increases, the standard deviation of the sample mean falls drastically. In sum, this data simulation reinforces the central limit theorem discussed in Sect. 8.6.

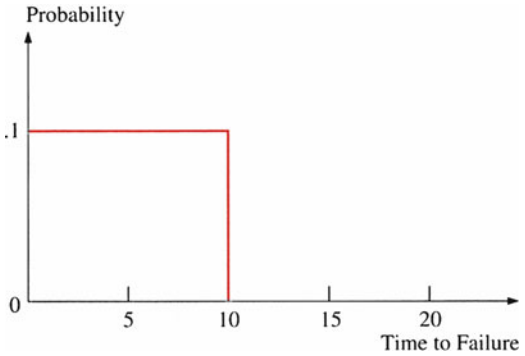


Fig. 8.9 Uniform distribution from 0 to 1

Table 8.13 Sample means for five different sample sizes

$n = 2$		$n = 5$		$n = 10$		$n = 25$		$n = 50$	
MTB > PRINT K1-K40		MTB > PRINT K1-K40		MTB > PRINT K1-K40		MTB > PRINT K1-K40		MTB > PRINT K1-K40	
K1	0.766039	K1	0.547407	K1	0.508664	K1	0.477483	K1	0.467134
K2	0.307122	K2	0.631557	K2	0.572561	K2	0.491987	K2	0.486903
K3	0.328305	K3	0.263628	K3	0.449149	K3	0.440873	K3	0.525870
K4	0.275898	K4	0.503923	K4	0.444761	K4	0.529234	K4	0.546720
K5	0.407282	K5	0.372527	K5	0.514358	K5	0.395570	K5	0.491633
K6	0.783204	K6	0.676701	K6	0.467277	K6	0.483190	K6	0.561301
K7	0.568790	K7	0.345473	K7	0.449323	K7	0.496404	K7	0.470143
K8	0.339255	K8	0.579194	K8	0.319622	K8	0.469850	K8	0.580831
K9	0.363014	K9	0.697247	K9	0.404897	K9	0.489090	K9	0.487812
K10	0.594667	K10	0.337961	K10	0.425953	K10	0.553480	K10	0.495207
K11	0.676628	K11	0.565731	K11	0.739907	K11	0.381792	K11	0.527002
K12	0.313933	K12	0.412871	K12	0.364374	K12	0.592549	K12	0.446332
K13	0.209292	K13	0.364198	K13	0.407843	K13	0.456389	K13	0.623809
K14	0.185163	K14	0.409698	K14	0.497028	K14	0.475342	K14	0.425805
K15	0.167203	K15	0.403446	K15	0.421893	K15	0.442597	K15	0.563418
K16	0.546706	K16	0.444514	K16	0.558162	K16	0.403973	K16	0.443062
K17	0.284282	K17	0.589411	K17	0.279482	K17	0.411229	K17	0.478683
K18	0.413619	K18	0.489910	K18	0.464799	K18	0.463406	K18	0.484854
K19	0.298615	K19	0.372339	K19	0.559232	K19	0.611943	K19	0.516529
K20	0.355693	K20	0.751555	K20	0.572471	K20	0.624390	K20	0.507905
K21	0.709642	K21	0.594944	K21	0.642801	K21	0.518310	K21	0.480531
K22	0.661339	K22	0.522054	K22	0.526308	K22	0.453991	K22	0.432330
K23	0.418546	K23	0.513973	K23	0.439639	K23	0.483019	K23	0.539987
K24	0.275210	K24	0.311244	K24	0.419171	K24	0.449016	K24	0.527544
K25	0.661684	K25	0.525532	K25	0.495180	K25	0.503539	K25	0.497368
K26	0.805315	K26	0.449068	K26	0.400065	K26	0.498799	K26	0.524032

K27	0.542858	K27	0.462357	K27	0.509161	K27	0.534755	K27	0.482675
K28	0.149722	K28	0.491276	K28	0.497628	K28	0.547716	K28	0.494472
K29	0.411248	K29	0.731628	K29	0.511615	K29	0.489730	K29	0.516521
K30	0.250607	K30	0.243674	K30	0.442089	K30	0.401362	K30	0.444785
K31	0.236916	K31	0.514128	K31	0.521398	K31	0.431294	K31	0.500245
K32	0.316803	K32	0.579376	K32	0.474808	K32	0.501472	K32	0.434281
K33	0.546881	K33	0.660470	K33	0.361446	K33	0.641570	K33	0.471526
K34	0.508401	K34	0.604544	K34	0.537571	K34	0.436615	K34	0.492110
K35	0.260588	K35	0.384347	K35	0.341068	K35	0.534256	K35	0.473473
K36	0.687794	K36	0.321559	K36	0.419929	K36	0.554134	K36	0.474246
K37	0.783227	K37	0.294595	K37	0.464406	K37	0.433804	K37	0.464606
K38	0.423624	K38	0.408459	K38	0.503744	K38	0.485380	K38	0.545458
K39	0.618296	K39	0.508032	K39	0.600979	K39	0.465627	K39	0.514041
K40	0.379861	K40	0.545842	K40	0.571873	K40	0.420188	K40	0.418817

```
MTB > GSTD
* NOTE  * Standard Graphics are enabled.
          Professional Graphics are disabled.
          Use the GPRO command to enable Professional Graphic.
MTB > HISTOGRAM C1;
SUBC> START=0.1;
SUBC> INCREMENT=0.05.
```

Character Histogram

```
Histogram of  C1    N = 40

Midpoint     Count
  0.1000         0
  0.1500         2    **
  0.2000         2    **
  0.2500         3    ***
  0.3000         7    *******
  0.3500         4    ****
  0.4000         6    ******
  0.4500         0
  0.5000         1    *
  0.5500         4    ****
  0.6000         2    **
  0.6500         2    **
  0.7000         3    ***
  0.7500         1    *
  0.8000         3    ***
```

Fig. 8.10 Histogram of 40 sample means ($n = 2$)

```
MTB > GSTD
* NOTE  * Standard Graphics are enabled.
         Professional Graphics are disabled.
         Use the GPRO command to enable Professional Graphic.
MTB > HISTOGRAM C2;
SUBC> START=0.1;
SUBC> INCREMENT=0.05.
```

Character Histogram

```
Histogram of  C2    N = 40

Midpoint      Count
  0.1000          0
  0.1500          0
  0.2000          0
  0.2500          2   **
  0.3000          3   ***
  0.3500          5   *****
  0.4000          5   *****
  0.4500          3   ***
  0.5000          7   *******
  0.5500          4   ****
  0.6000          5   *****
  0.6500          2   **
  0.7000          2   **
  0.7500          2   **
```

Fig. 8.11 Histogram of 40 sample means ($n = 5$)

```
MTB > GSTD
* NOTE  * Standard Graphics are enabled.
          Professional Graphics are disabled.
          Use the GPRO command to enable Professional Graphics.
MTB > HISTOGRAM C3;
SUBC> START=0.1;
SUBC> INCREMENT=0.05.
```

Character Histogram

```
Histogram of  C3    N = 40

Midpoint      Count
  0.1000          0
  0.1500          0
  0.2000          0
  0.2500          0
  0.3000          2   **
  0.3500          3   ***
  0.4000          6   ******
  0.4500         10   **********
  0.5000          9   *********
  0.5500          7   *******
  0.6000          1   *
  0.6500          1   *
  0.7000          0
  0.7500          1   *
```

Fig. 8.12 Histogram of 40 sample means ($n = 10$). Histogram of 40 sample means ($n = 5$)

```
MTB > GSTD
* NOTE  * Standard Graphics are enabled.
          Professional Graphics are disabled
          Use the GPRO command to enable Professional Graphics.
MTB > HISTOGRAM C4;
SUBC> START=0.1;
SUBC> INCREMENT=0.05.
```

Character Histogram

```
Histogram of  C4    N = 40

Midpoint      Count
  0.1000          0
  0.1500          0
  0.2000          0
  0.2500          0
  0.3500          0
  0.4000          0
  0.4500         11   ***********
  0.5000         12   ************
  0.5500          6   ******
  0.6000          3   ***
  0.6500          1   *
```

Fig. 8.13 Histogram of 40 sample means ($n = 25$). Histogram of 40 sample means ($n = 5$)

```
MTB > GSTD
* NOTE  * Standard Graphics are enabled.
         Professional Graphics are disabled.
         Use the GPRO command to enable Professional Graphics.
MTB > HISTOGRAM C5;
SUBC> START=0.1;
SUBC> INCREMENT=0.05.
```

Character Histogram

```
Histogram of  C5    N = 40

Midpoint      Count
  0.1000          0
  0.1500          0
  0.2000          0
  0.2500          0
  0.3000          0
  0.3500          0
  0.4000          1   *
  0.4500         13   *************
  0.5000         16   ****************
  0.5500          8   ********
  0.6000          2   **
```

Fig. 8.14 Histogram of 40 sample means ($n = 50$)

Chapter 9
Other Continuous Distributions and Moments for Distributions

Chapter Outline

Key Terms

Uniform distribution	Exponential distribution
Student's *t* distribution	Moments
Degree of freedom	Coefficient of variation
Chi-square distribution	Coefficient of skewness
F distribution	Coefficient of kurtosis
F variable	Noncentral chi-square distribution

C.-F. Lee et al., *Statistics for Business and Financial Economics,*
DOI 10.1007/978-1-4614-5897-5_9,

9.1 Introduction

Two very useful continuous distributions, the normal and lognormal distributions, were discussed in Chap. 7. Because many random variables have distributions that are not normal, in this chapter, we explore five other important continuous distributions and their applications. These five distributions are the uniform distribution, Student's t distribution, the chi-square distribution, the F distribution, and the exponential distribution. All are directly or indirectly used in analyzing business and economic data. The relationship between moments and distributions is also discussed in this chapter. Finally, we explore business applications of statistical distributions in terms of the first four moments for stock rates of return.

9.2 The Uniform Distribution

The simplest continuous probability distribution is called the *uniform distribution*. This probability distribution provides a model for continuous random variables that are evenly (or randomly) distributed over a certain interval. To picture this distribution, assume that the random variable X can take on any value in the range from, for example, 5 to 15, as indicated in Fig. 9.1. In a uniform distribution, the probability that the variable will assume a value within a given interval is proportional to the length of the interval. For example, the probability that X will assume a value in the range from 6 to 8 is the same as the probability that it will assume a value in the range from 9 to 11, because these two intervals are equal in length.

The uniform distribution has the following probability density function:

$$f(X) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq X \leq b \\ 0 & \text{elsewhere} \end{cases} \tag{9.1}$$

If the foregoing condition holds, then X is uniformly distributed, and the shape under the density function forms a rectangle, as shown in Fig. 9.1. The rectangle's area is equal to 1, which means that X is sure to take on some value between $a = 5$ and $b = 15$. Mathematically, we can express this as $P(5 \leq X \leq 15) = 1$.

Figure 9.1 shows a density function for a set of values between a and b. Each density is a horizontal line segment with constant height $1/(b - a)$ over the interval from a to b. Outside the interval, $f(X) = 0$. This means that for a uniformly distributed random variable X, values below a and values above b are impossible. Substituting $b = 15$ and $a = 5$ into Eq. 9.1, we obtain $1/(b - a) = 1/(15 - 5) = .1$, as indicated in Fig. 9.1.

From Chaps. 5 and 7, we know that the probability that X will fall below a point is provided by the area under the density curve and to the left of that point. In other words, the cumulative probability distribution function, $P(X \leq x) = (x - a)/(b - a)$, is represented by this area. The cumulative function for values of X

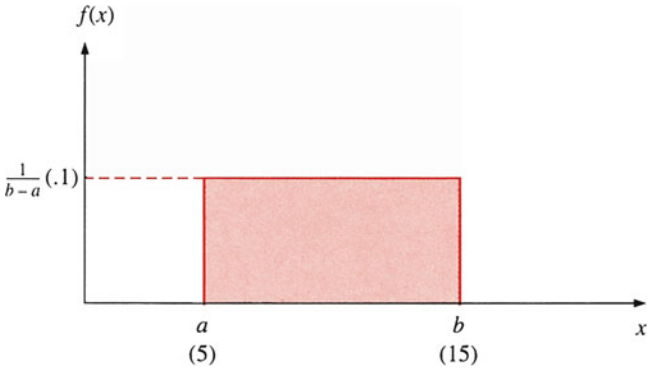


Fig. 9.1 The uniform probability distribution

between a and b is the area of the rectangle, which, again, is found by multiplying the height, $1/(b-a)$, times the base, $x-a$. To the left of a, the cumulative probabilities must be zero, whereas the probability that X lies "below points beyond b" must be 1.

The cumulative probabilities for a uniform distribution are

$$P(X \le x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \le x \le b \\ 1 & \text{if } x > b \end{cases} \tag{9.2}$$

Figure 9.2 shows the cumulative distribution function in terms of data indicated in Fig. 9.1. It presents the cumulative probabilities for $X = 5, X = 10, X = 15$, and $X = 20$ at points A, B, C, and D, respectively. Cumulative probabilities for these three points can be calculated as follows:

At point A: $P(X \le 5) = \dfrac{5-5}{15-5} = 0$

At point B: $P(X \le 10) = \dfrac{10-5}{15-5} = \dfrac{1}{2}$

At point C: $P(X \le 15) = \dfrac{15-5}{15-5} = 1$

At point D: $P(X \le 20) = P(X \le 15) + P(15 \le X \le 20) = 1 + 0 = 1$

The mean and standard deviation of a uniform distribution (see Appendix 1) can be shown as

$$\mu = E(X) = \frac{a+b}{2}$$
$$\sigma_X = \frac{b-a}{\sqrt{12}} \tag{9.3}$$

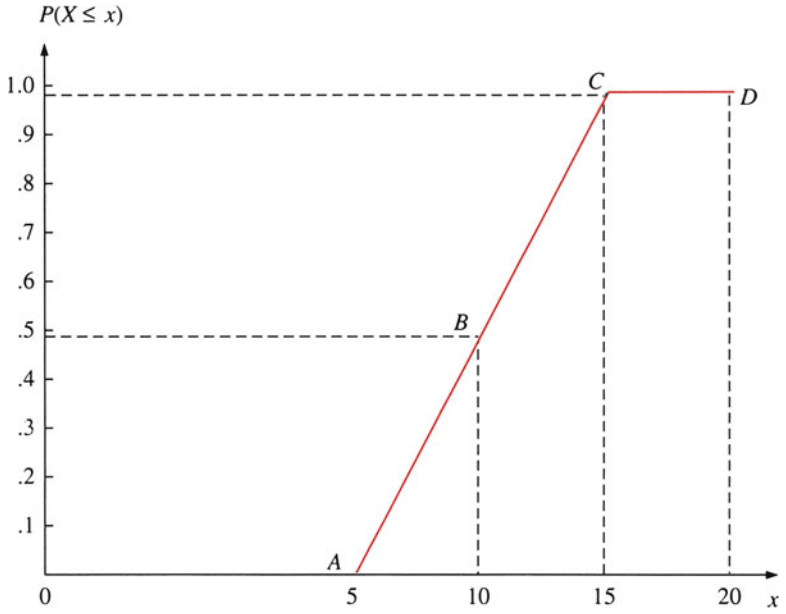


Fig. 9.2 Cumulative distribution function for the data of Fig. 9.1

Example 9.1 An Application of the Uniform Distribution in Quality Control. A quality control inspector for Gonsalves Company, which manufactures aluminum water pipes, believes that the product has varying lengths. Suppose the pipes turned out by one of the production lines of Gonsalves Company can be modeled by a uniform probability distribution over the interval 29.50–30.05 ft. The mean and standard deviation of X, the length of the aluminum water pipe, can be calculated as follows. Substituting $b = 30.05$ ft and $a = 29.50$ ft in Eq. 9.3, we obtain

$$\mu = \frac{30.05 + 29.50}{2} = 29.775 \text{ ft}$$

and

$$\sigma_X = \frac{30.05 - 29.50}{\sqrt{12}} = .1588 \text{ ft}$$

This information can be used to create a control chart to determine whether the quality of the water pipes is acceptable. The control chart and its use in statistical quality control will be discussed in Chap. 10.

Computer simulation is an application of statistics that frequently relies on the uniform distribution. In fact, the uniform distribution is the underlying mechanism for this often-complex procedure. Thus, although not so many "real-world" populations resemble this distribution as resemble the normal, the uniform

distribution is important in applied statistics. For example, managers may use the uniform distribution in a simulation model to help them decide whether the company should undertake production of a new product.[1] Basic concepts of investment decision making can be found in Sect. 21.8.

9.3 Student's *t* Distribution

Student's *t* distribution was first derived by W. S. Gosset in 1908. Because Gosset wrote under the pseudonym "A Student," this distribution became known as Student's *t* distribution.

If the sampled population is normally distributed with mean μ and variance σ_X^2, the sample size n is equal to or larger than 30, and σ_X^2 is known, then from the last chapter, we know that the Z score for sample mean $\bar{X}$ defined as

$$Z = \frac{\bar{X} - \mu}{\sigma_X / \sqrt{n}} \tag{8.7}$$

which we met as Eq. 8.7, has a normal distribution with mean 0 and variance 1. Under most circumstances, however, the population variance is not known. In order for us to conduct various types of statistical analysis, we need to know what happens to Eq. 8.7 when we replace the population standard deviation σ_X by the sample standard deviation s_X. We then have the following equation for the t statistic:

$$t = \frac{\bar{X} - \mu}{s_X / \sqrt{n}} \tag{9.4}$$

Thus, the Z of Eq. 8.7 has only one source of variation: each sample has a different $\bar{X}$. Equation 9.4, however, has two sources of variation: both the sample mean $\bar{X}$ and the sample standard deviation s_X change from sample to sample. Thus, the term on the right-hand side of Eq. 9.4 follows a sampling distribution different from the normal distribution, which is the distribution followed by the term on the right-hand side of Eq. 8.7. Equation 9.4 is used only when the population from which the n sample items are drawn is normally distributed and the sample size (n) is smaller than 30.

The t distribution forms a family of distributions that are dependent on a parameter known as the *degrees of freedom*. For the t variable in Eq. 9.4, the degrees of freedom (v) are $(n - 1)$, where n is the sample size. In general, the degrees of freedom for a t statistic are the degrees of freedom associated with the sum of squares used to obtain an estimate of the variance. The variance estimate

[1] See Lee C.F.: Financial Analysis and Planning: Theory and Application, pp. 358–363. Reading, Addison-Wesley (1985)

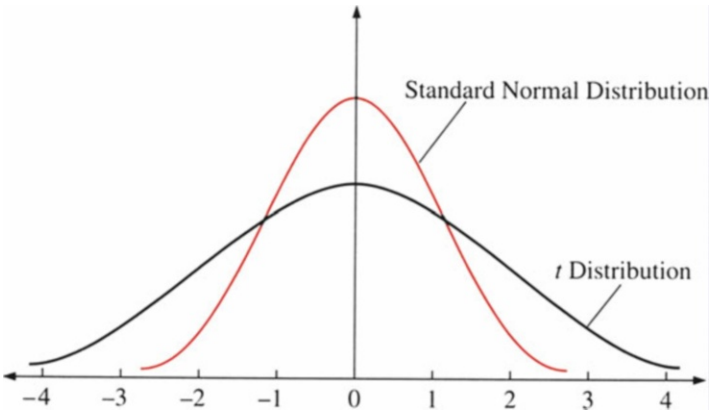


Fig. 9.3 The *t* distribution and the standard normal distribution

depends not only on the size of sample but also on how many parameters must be estimated with the sample. The more data we have, the more confidence we can have in our results; the more parameters we have to estimate, the less confidence we have. Statisticians keep track of these two factors by calculating the degrees of freedom as follows:

$$\begin{array}{c}\text{Degrees of}\\ \text{freedom}\end{array} = \begin{array}{c}\text{number of}\\ \text{observations}\end{array} - \begin{array}{c}\text{number of parameters that must}\\ \text{be estimated beforehand}\end{array}$$

Here we calculate s_X by using n observations and estimating one parameter (the mean). Thus, there are $(n - 1)$ degrees of freedom.

The *t* distribution is a symmetric distribution with mean 0. Its graph is similar to that of the standard normal distribution, as Fig. 9.3 shows. However, the tail areas are greater for the *t* distribution, and the standard normal distribution is higher in the middle. The larger the number of degrees of freedom, the more closely the *t* distribution resembles the standard normal distribution. As the number of degrees of freedom increases without limit, the *t* distribution approaches the standard normal distribution. In fact, the standard normal distribution *is* a *t* distribution with an infinite number of degrees of freedom.

To determine whether the normal distribution or the Student's *t* distribution is more suitable for describing stocks' rates of return, Blattberg and Gonedes (1975, *Journal of Business*, pp. 244–280) used both daily and weekly stock rates of return for Dow Jones 30 companies to estimate the degrees of freedom for these two kinds of rates of return. They found, for example, that the degrees of freedom for Allied Chemical are 5.04 when daily data is used and 89.98 when weekly data is used. This indicates that the student's *t* distribution is more suitable for daily data for Allied Chemical, whereas the normal distribution better describes weekly data for Allied Chemical.

In addition, they found that the average degree of freedom for daily rates of return for these 30 companies is 4.79. The average degree of freedom in terms of

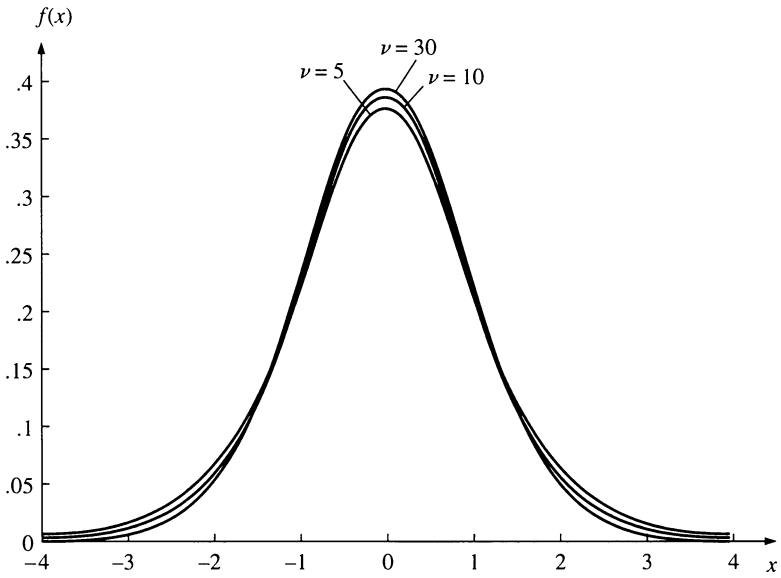


Fig. 9.4 *t* distributions of three different degrees of freedom

weekly rate of return for these 30 companies is 11.22. They concluded that Student's *t* distribution is more suitable for describing daily stock rate of return distribution, and normal distribution is more suitable for weekly rate of return distribution. Hence, *t* distribution is an important distribution for describing daily stock rate of return.

The *t* table, as presented in Table A4 at the end of the book, gives the value, t_α, such that the probability of the *t* value larger than t_α is equal to α. The percentage cutoff point t_α is defined as that point at which

$$P(t > t_\alpha) = \alpha \tag{9.5}$$

Because the distribution is symmetric around 0, only positive *t* values (upper-tail areas) are tabulated. The lower α cutoff point is $-t_\alpha$, because

$$P(t < -t_\alpha) = P(t > t_\alpha) = \alpha \tag{9.6}$$

In general, we denote a cutoff point for *t* by $t_{\alpha,v}$ where α is the probability level and *v* is the degrees of freedom. The number of degrees of freedom determines the shape of the *t* distribution. Figure 9.4 shows *t* distributions of varying degrees of freedom.

Example 9.2 Using the t *Distribution to Analyze Audit Sampling Information.* Let's borrow information presented in Sect. 8.7 to see how the *t* distribution can be used to do audit sampling analysis.

The sample mean and the sample variance for 30 trade accounts receivable balances are

$$\bar{X} = \$202.10 \quad \text{and} \quad s_X^2 = \$719.164$$

From Table A4, we know that the t statistics with $30 - 1 = 29^{\circ}$ of freedom and $\alpha = .05$ is 1.6991. Substituting related information into Eq. 9.4, we obtain

$$1.699 = \frac{\$202.10 - \mu}{\sqrt{719.164/30}} = \frac{\$202.10 - \mu}{4.896}$$

This implies that there is a 5 % chance that the average population account receivable value will be smaller than $\$202.10 - \$(11.699)(4.896) = \$193.78$. By symmetry, there is also a 5 % chance that the average population account receivable value will be larger than $\$202.10 + \$(1.699)(4.896) = \$210.42$.

Other applications of the t distribution appear in Chaps. 10 and 11, and we will encounter more when we discuss regression analysis in Chaps. 13, 14, 15, and 16.

9.4 The Chi-Square Distribution and the Distribution of Sample Variance

In this section, we first show how a chi-square distribution can be derived from a standard normal distribution and then derive the distribution of a sample variance.

9.4.1 The Chi-Square Distribution

The *chi-square distribution* (χ^2) is a continuous distribution ordinarily derived as the sampling distribution of a sum of squares of independent standard normal variables. For instance, let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a normal distribution with mean μ and variance σ_X^2. Because these variables are not standardized, we can standardize them as

$$Z_i = \frac{X_i - \mu}{\sigma_X}$$

where Z_i is normally distributed with mean 0 and variance 1.

Now, if we define a new variable Y such that

$$Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma_X} \right)^2 \tag{9.7}$$

it can be shown that this new variable is distributed as χ^2 with n degrees of freedom.[2]

Equation 9.6 can be rewritten as[3]

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma_X^2} = \frac{n(\bar{X} - \mu)^2}{\sigma_X^2} + \frac{(n-1)s_X^2}{\sigma_X^2} \tag{9.8}$$

where

$$s_X^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}; \quad \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma_X^2}$$

has an x^2 distribution with n degrees of freedom, as discussed in Eq. 9.6. In addition, from the last chapter, we know that $\bar{X}$ is normally distributed with mean μ and variance σ_X^2/n, so $\sqrt{n}(\bar{X} - \mu)/\sigma_X$ is normally distributed with mean 0 and variance 1. It can be shown that $n(\bar{X} - \mu)^2/\sigma_X^2$ has an x^2 distribution with 1° of freedom. From this information, it can be proved that

$$\frac{(n-1)s_X^2}{\sigma_X^2}$$

defined in Eq. 9.8, has a χ^2 distribution with $(n - 1)$ degrees of freedom.[4]

---

[2] First, it can be proved that $(X_i - \mu)^2/\sigma_X^2$ is a χ^2 distribution with I degree of freedom. Then, by using the additive property of x^2 distribution, we can prove that $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma_X}\right)^2$ is also a χ^2 distribution with n degrees of freedom.

[3] Since

$$\begin{aligned}\sum_{i=1}^{n} (X_i - \mu)^2 &= \sum_{i=1}^{n} (X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \sum_{i=1}^{n} (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \sum_{i=1}^{n} (X_i - \mu) + \sum_{i=1}^{n} (\bar{X} - \mu)^2 \\ &= \sum_{i=1}^{n} (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2\end{aligned}$$

because

$$2(\bar{X} - \mu) \sum_{i=1}^{n} (X_i - \bar{X}) = 0 \tag{9.A}$$

by dividing Eq. 9.A by σ_X^2, we obtain Eq. 9.8.

[4] In addition to the condition described here, it is also necessary to assume that $\bar{X}$ is independent of s_X^2.

$$\frac{(n-1)s_X^2}{\sigma_X^2}$$

can be redefined as expressed in Eq. 9.9:

$$\frac{(n-1)s_X^2}{\sigma_X^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_X}\right)^2 \tag{9.9}$$

where s_X^2 and σ_X^2 are sample variance and population variance, respectively. The left-hand side of Eq. 9.9 implies that the ratio of sample variance to population variance, multiplied by $(n-1)$, has a χ^2 distribution with $(n-1)$ degrees of freedom. The χ^2 distribution defined in Eq. 9.9 can be used to describe the distribution of s_X^2, which will be discussed later in this section.

The χ^2 distribution is a skewed distribution, and only nonnegative values of the variable χ^2 are possible. It depends on a single parameter, the degrees of freedom $v = n - 1$. The χ^2 distributions for degrees of freedom 5, 10, and 30 are graphed in Fig. 9.5. The figure shows that the skewness decreases as the degrees of freedom increase. In fact, as the degrees of freedom increase to infinity, the χ^2 distribution approaches a normal distribution.[5]

Critical values of the χ^2 distributions are given in Table A5 in Appendix A.[6] They are defined by

$$P\left(\chi^2 \geq \chi^2_{\alpha,v}\right) = \alpha \tag{9.10}$$

where is that value for the χ^2 distribution with v degrees of freedom such that the area to the right (the probability of a larger value) is equal to α. For example, the upper 5 % point for χ^2 with 10 degree of freedom, $\chi^2_{.05,10}$, is 18.307 (see Fig. 9.6 and Table A5). In other words, $P(\chi^2 > 18.307) = .05$. In addition, $P(\chi^2 < 18.307) = 1 - .05 = .95$.

The mean and variance of this distribution are equal to the number of degrees of freedom and twice the number of degrees of freedom. That is,

[5] Johnson, W. L., Katz S.: In Continuous Univariate Distribution I, pp. 170–181. Houghton Mifflin, Boston, 1970, show that a normalized x^2 distribution approaches a standard normal distribution when the number of degrees of freedom approaches infinity. The normalized statistic is defined as $(\chi_v^2 - v)/\sqrt{2v}$.

[6] We can approximate ${x^2}_\alpha$ by the formula

$$\chi^2_\alpha = v\left(1 - \frac{2}{9v} + z_\alpha\sqrt{\frac{2}{9v}}\right)^3$$

where v = degrees of freedom and z_α = standard normal value (from Table A.3).

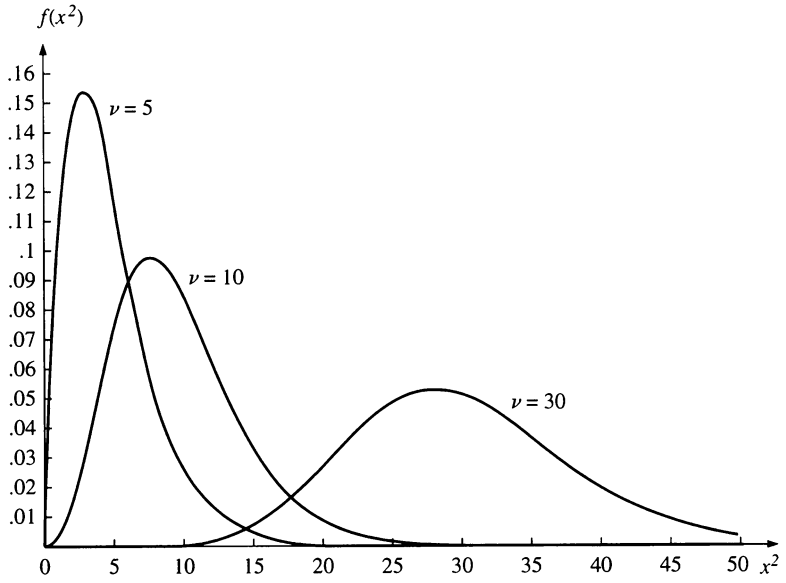


Fig. 9.5 The χ^2 distributions with three different degrees of freedom

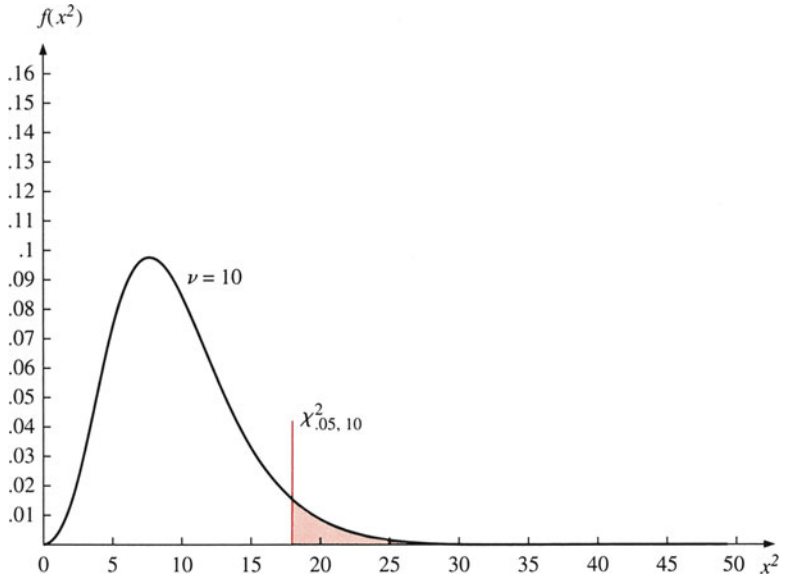


Fig. 9.6 The χ^2 distribution with 10° of freedom

$$E(\chi_v^2) = v, \quad \text{and} \quad \text{Var}(\chi_v^2) = 2v \tag{9.11}$$

where v is the degree of freedom of a χ^2 distribution.

9.4.2 The Distribution of Sample Variance

The properties of the χ^2 distribution can be used to find the mean and variance of the sampling distribution of the sample variance (s_X^2).

9.4.2.1 The Mean of s_X^2

From the definition of the mean for a χ^2 distribution, we obtain

$$E\left[\frac{(n-1)s_X^2}{\sigma_X^2}\right] = n - 1$$

Because $E(a \cdot X) = a \cdot E(X)$, we have

$$\frac{(n-1)}{\sigma_X^2} E\left(s_X^2\right) = n - 1$$

Thus,[7]

$$E\left(s_X^2\right) = \sigma_X^2 \tag{9.12}$$

Equation 9.12 implies that the mean of the sample variance is equal to the population variance.

9.4.2.2 The Variance of s_X^2

On the basis of the definition of the variance for a χ^2 distribution, we have

$$\mathrm{Var}\left[\frac{(n-1)S_X^2}{\sigma_X^2}\right] = 2(n-1)$$

Because $\mathrm{Var}(aX) = a^2 \cdot \mathrm{Var}(X)$, we have

$$\frac{(n-1)^2}{\sigma_X^4} \mathrm{Var}\left(s_X^2\right) = 2(n-1)$$

so

$$\mathrm{Var}\left(s_X^2\right) = \frac{2\sigma_X^4}{n-1} \tag{9.13}$$

[7] This result suggests why $\sum_{i=1}^{n} (X_i - \bar{X})^2 / n - 1$ instead of $\sum_{i=1}^{n} (X_i - \bar{X})^2 / n$ is an unbiased estimator for the population variance, σ_X^2. Unbiased estimators will be discussed in Chap. 10.

This is the variance of the sample variance. In sum, if X is normally distributed, then the mean and variance of s_X^2 are σ_X^2 and $2\sigma_X^4/(n-1)$, respectively. We will explore applications of the χ^2 distribution and the distribution of sample variance in Chaps. 10 and 11 when we discuss confidence intervals and hypothesis testing for population variances.

Drawing on the concepts of the χ^2 distribution and the normal distribution, we can interpret the t distribution by rewriting Eq. 9.4′ as

$$t = \frac{(\bar{X} - \mu)/(\sigma_X/\sqrt{n})}{s_X/\sigma_X} \tag{9.4′}$$

In Eq. 9.4′, $(\bar{X} - \mu)/(\sigma_X/\sqrt{n})$ is normally distributed with mean 0 and variance 1; it is a standard normal distribution. S_x/σ_x is a square root of a χ^2-distributed variable with $(n-1)$ degrees of freedom divided by $v = n - 1$. Hence, a t distribution with v degrees of freedom is the ratio between a standard normal variable and a transformed χ^2 variable:

$$t_v = \frac{Z}{\sqrt{\chi_v^2/v}} \tag{9.14}$$

9.5 The *F* Distribution

Some problems revolve around the value of a single population variance, but often it is a comparison of the variances of two populations that is of interest. This will be discussed in Chaps. 13, 14, and 15. In addition, we may want to know whether the means of three or more populations are equal. This will be discussed in Chap. 12. The *F distribution* is used to make inferences about these kinds of issues.

Assume two populations, each having a normal distribution. We draw two independent random samples with sample sizes n_X and n_Y and population variances σ_X^2 and σ_Y^2. From each sample, we can compute sample variances S_X^2 and S_Y^2. Then, the random variable of Eq. 9.15 follows a distribution known as the F distribution:

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \tag{9.15}$$

Equation 9.15 can be rewritten as

$$F = \frac{\chi_{v_1}^2(X)/(n_X - 1)}{\chi_{v_2}^2(Y)/(n_Y - 1)} \tag{9.14′}$$

where $\chi_{v_1}^2(X) = (n_X - 1)S_X^2/\sigma_X^2$ and $\chi_{v_2}^2(Y) = (n_Y - 1)S_Y^2/\sigma_Y^2$; $v_1 = n_x - 1$; $v_2 = n_y - 1$.

In other words, a random variable formed by the ratio of two independent chi-square variables, each divided by its degrees of freedom, is called an *F variable*.

The *F* distribution has an asymmetric probability density function defined only for nonnegative values. It should be observed that the *F* distribution is completely determined by two parameters, v_1 and v_2, which are degrees of freedom. These density functions with different sets of degrees of freedom are illustrated in Fig. 9.7.

The cutoff points $F_{v_1,v_2,\alpha}$, for α equal to .05, .025, .01, and .005, are provided in Table A6 at the end of this book. For example, in the case of 10 numerator degrees of freedom and six denominator degrees of freedom,

$$F_{10,6,.05} = 4.06 \quad F_{10,6,.025} = 5.46$$
$$F_{10,6,.01} = 7.87 \quad F_{10,6,.005} = 10.25$$

MINITAB output for $F_{10,\,6}$ is presented in Fig. 9.8. Hence,

$$P(F_{10.6} > 4.06) = .05 \qquad P(F_{10.6} > 5.46) = .025$$
$$P(F_{10.6} > 7.87) = .01 \qquad P(F_{10.6} > 10.25) = .005$$

These probabilities also can be calculated by using MINITAB as shown here.

```
MTB > SET C1
DATA> 4.06 5.46 7.87 10.25
DATA> END
MTB > CDF C1;
SUB > F 10 6.
```

Cumulative Distribution Function

```
F distribution with 10 DF in numerator and 6 DF in
denominator
  x P( X <= x)
  4.0600 0.9500
  5.4600 0.9750
  7.8700 0.9900
  10.2500 0.9950
```

By subtracting 1 from .95, we obtain .05; by subtracting 1 from .975, we obtain .025; by subtracting 1 from .99, we obtain .01; finally, by subtracting 1 from .9950, we obtain .005. In practice, we usually place the larger sample variance in the numerator. The four significance levels listed here are the cutoff points that are often used to test the hypothesis of equality of population variances, which will be discussed in Chaps. 11 and 12. When the population variances are equal, Eq. 9.15 becomes

$$F = \frac{S_X^2}{S_Y^2} \tag{9.16}$$

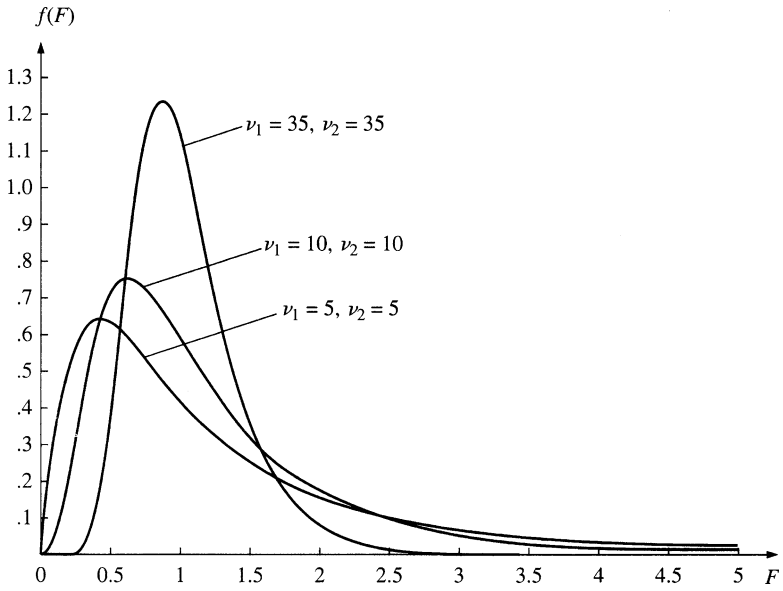


Fig. 9.7 F distributions with three different sets of degrees of freedom

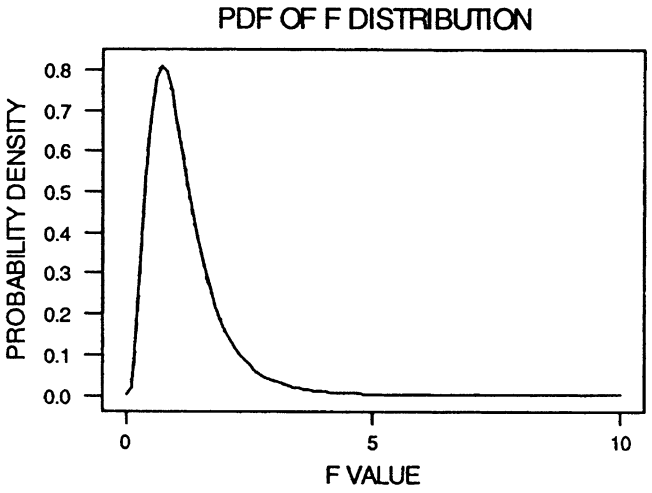


Fig. 9.8 MINITAB output for $F_{10,\,6}$

The right-hand side of Eq. 9.16 is the ratio of two sample variances.

Applications of the F distribution will be discussed in Chaps. 11 and 12 and in the chapters related to regression analysis.

9.6 The Exponential Distribution (Optional)

The *exponential distribution* is related to the Poisson distribution, which, as we noted in Chap. 6, is often applied to occurrences of an event over time. The Poisson distribution is the distribution of the number of occurrences of an event in a given time interval of length t. The single parameter of the Poisson distribution is λ, the intensity of the process. Think of the number as the average occurrence of the event being counted. For example, say, the average arrival rate of customers at the Brownell Bank is 5 per 100 s. Suppose that instead of the number of occurrences in a given time period, we are interested in the amount of time until the first customer arrives at the bank. This is a problem to be solved by the exponential distribution instead of the Poisson distribution. As another example, if the number of traffic accidents in an interval of time follows the Poisson distribution, the length of time from one accident to another follows the exponential distribution. The exponential distribution can also be applied to (1) the length of time that must pass before the first incoming telephone call and (2) the length of time someone must wait for a cab in a given location, such as Penn Station in New York City.

Denoting the mean rate at which events occur over time by λ and denoting the time until the first event occurs by t, we can use the Poisson probability density function to derive the exponential probability density function (PDF).[8] It is

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t}, \quad t \geq 0 \\ &= 0, \qquad t < 0 \end{aligned} \tag{9.17}$$

where $\lambda > 0$ is the only parameter.

From Eq. 9.38 we know that the cumulative probability function is given by

$$\begin{aligned} F(t) = P(T \leq t) &= 1 - e^{-\lambda t}, \quad t \geq 0 \\ &= 0, \qquad t < 0 \end{aligned} \tag{9.18}$$

where T is a random variable representing time and t is a specific value.

Figure 9.9 represents four exponential functions for which λ equals 3, 2, 1, and $\frac{1}{2}$. From Appendix 2, we know that

$$E(T) = \frac{1}{\lambda} \tag{9.19}$$

$$\text{Var}(T) = \frac{1}{\lambda^2} \tag{9.20}$$

[8] See Appendix 2 for the derivation.

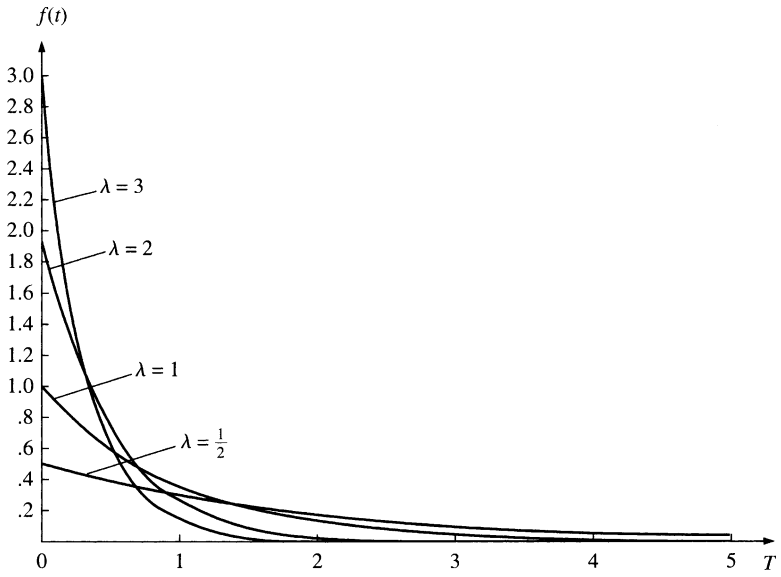


Fig. 9.9 Four exponential density functions specified by four alternative values of λ

Example 9.3 "No More Than 8 Items in This Line, Please!". Under fairly plausible assumptions about the behavior of clerks at supermarket check-out counters, it is possible to show that the time T (in minutes) a customer spends at a check-out counter is a random variable with the exponential distribution described by Eq. 9.17.

Suppose a supermarket check-out counter has a mean number of customers per minute $= \frac{1}{3}$; that is, $\lambda = \frac{1}{3}$. Our task is to find the probability that the length of time between a pair of customer arrivals is less than 6 min.

Substituting $\lambda = \frac{1}{3}$ and $t = 6$ into Eq. 9.18, we obtain $F(T \leq 6) = 1 - e^{-6/3}$. And referring to Table A7 of Appendix A (or to a hand calculator), we find $P(T < 6) = 1 - .1353 = .8647$. Thus, the probability that the service time available between two customer arrivals at the check-out counter will be less than 6 min is approximately .86. Alternatively, the probability .8647 can be obtained by MINITAB as shown here:

```
MTB > CDF 6;
SUBC> EXPONENTIAL 3.
Cumulative Distribution Function
Exponential with mean = 3.00000
x              P(X <= x)
6.0000           0.8647
```

9.7 Moments and Distributions (Optional)

The properties of a distribution can be described in many ways, but the most popular approach is by means of a set of measurements called moments. *Moments* describe the central tendency, degree of dispersion, asymmetry, peakedness, and many other aspects of a distribution. This section discusses only the first four moments of a distribution; they are the most important statistical characteristics.

The first k moments can be defined either as

$$\mu'_k = E\left(X^k\right) \tag{9.21}$$

or

$$\mu_k = E\left[(X-\mu)^k\right] \tag{9.22}$$

Equation 9.21 defines the k moments about the origin, and Eq. 9.22 defines the moments about the population mean μ. (The relationship between μ'_k and μ'_k is discussed in Appendix 3.) The *population mean* is the first moment about the origin. We obtain the first moment of a distribution about the origin by letting $k = 1$ in Eq. 9.21. It is defined as follows:

$$\mu'_1 = E(X) = \mu$$

This is the population mean of X. Following Eq. 4.1, we can define μ for a discrete variable as

$$\mu = \sum_{i=1}^{N} X_i / N \tag{4.2}$$

where N is the total number of observations in the population. The sample mean $\bar{X}$ associated with μ can be defined as

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{4.1}$$

where n is the sample size.

9.7.1 The Second Moment and the Coefficient of Variation

The second moment about the mean, the *variance*, is a measure of the dispersion of the random variable around the mean. The larger the variance, the more dispersed the distribution. Letting $k = 2$ in Eq. 9.22, we obtain

$$\mu_2 = \sigma_X^2 = E[X - E(X)]^2$$

This is the population variance of X. Following Eq. 4.5, we can define the population variance for a discrete variable as

$$\sigma_X^2 = \sum_{i=1}^{N} (X_i - \mu)^2 / N \tag{4.5}$$

The sample variance (s_X^2) associated with X can be defined as

$$s_X^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1) \tag{4.7}$$

Following Eq. 4.12, we can define the sample *coefficient of variation* (CV) as

$$\text{CV} = \frac{s_X}{\bar{X}} \tag{4.12}$$

9.7.2 The Third Moment and the Coefficient of Skewness

The third moment about the mean – *skewness*, which characterizes the asymmetry of the distribution – is given by

$$\mu_3 = E[X - E(X)]^3$$

Following Eq. 4.15, we can define the population skewness for a discrete variable as

$$\mu_3 = \sum_{i=1}^{N} (X_i - \mu)^3 / N \tag{4.15}$$

Following Eq. 4.16, we can define the *coefficient of skewness* (CS), which is a relative measure of asymmetry, as

$$\text{CS} = \frac{\mu_3}{\sigma^3} \tag{4.16}$$

Following Eq. 4.16a, we can define the sample coefficient of skewness (SCS) as

$$\text{SCS} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3 / (n-1)}{s_X^3} \tag{4.16a}$$

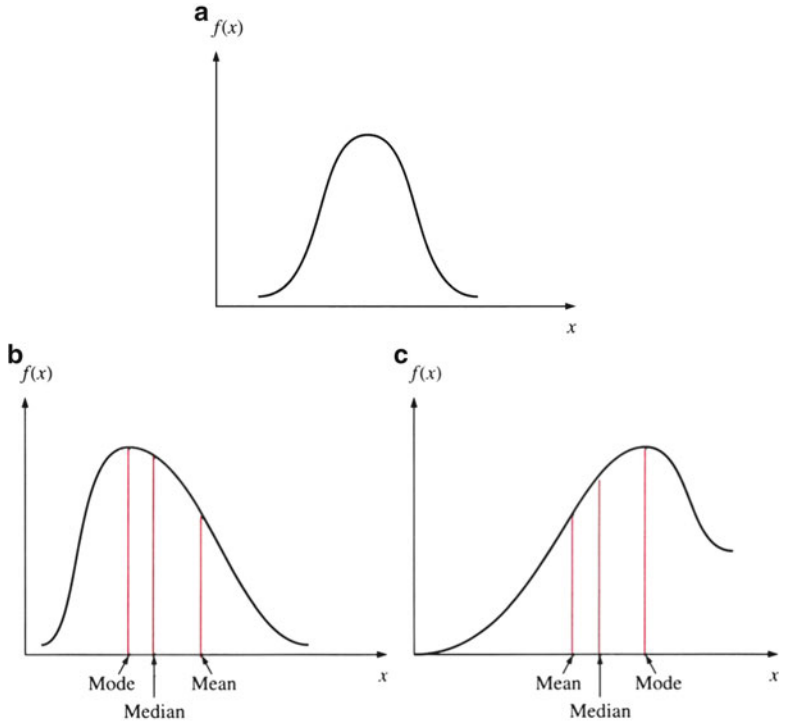


Fig. 9.10 (**a**) Zero skewness, (**b**) Positive skewness, and (**c**) negative skewness

where

$$s_X^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$$

Figures 9.10a, b, and c present graphs of distributions with differing degrees of symmetry. Figure 9.10a shows a symmetrical distribution – that is, a distribution with zero skewness. Note that the symmetrical distribution's measures of central tendency (the mean, median, and mode) all coincide. We can also see that the half of the distribution above the mode is a mirror image of the half of the distribution below the mode.

Figure 9.10b presents a distribution that is said to be positively skewed because the distribution tapers off more slowly to the right of the mode than to the left. It is clear that the mean, median, and mode do not coincide. Here, the mode is smaller than the median and the mean.

Figure 9.10c presents a distribution that is said to be negatively skewed because the distribution tapers off more slowly to the left of the mode than to the right. Once again, the mean, median, and mode do not coincide. Here the median and mean lie to the left of the mode.

9.7.3 Kurtosis and the Coefficient of Kurtosis

The fourth moment about the mean – *kurtosis*, which characterizes the degree of peakedness – is defined by

$$\mu_4 = E[X - E(X)]^4$$

For discrete variables, the population kurtosis can be defined as

$$\mu_4 = \sum_{i=1}^{N} (X_i - \mu)^4 / N \tag{9.23}$$

and can be estimated in terms of sample data as follows:

$$\text{Sample kurtosis} = \sum_{i=1}^{n} (X_i - \bar{X})^4 / n$$

The relative peakedness of a distribution is expressed by the ratio of the fourth moment to the square of the second moment. It is called *coefficient of kurtosis* (CK):

$$\text{CK} = \mu_4 / \mu_2^2 \tag{9.24}$$

This ratio measures the degree of peakedness relative to the level of dispersion. Using sample information, we can estimate the coefficient of kurtosis by

$$\text{SCK} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4 / (n-1)}{s_X^4} \tag{9.25}$$

Of two distributions having the same dispersion, the one with the larger kurtosis ratio has more observations concentrated near the mean and also at the tails of the distribution (at the expense of the intermediate area).

9.7.4 Skewness and Kurtosis for Normal and Lognormal Distributions

The bell-shaped normal curve is characterized by the *mesokurtic* shape: a value of 3 for the coefficient of kurtosis as defined in Eq. 9.25. Distributions with values of the kurtosis ratio greater than 3 are *leptokurtic*. These distributions are more peaked than the standard mesokurtic (normal curve) shape. Distributions with values of the

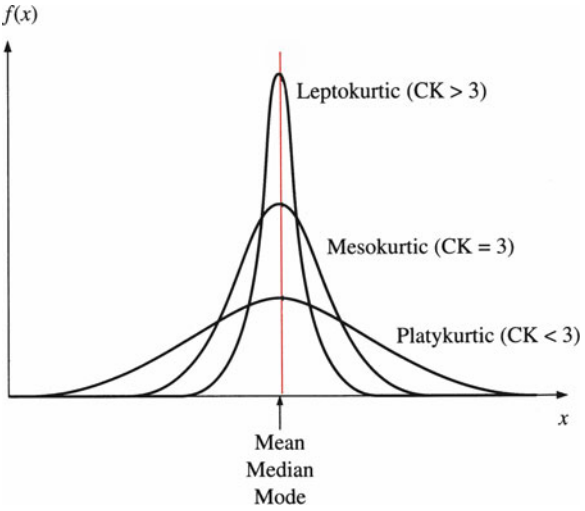


Fig. 9.11 Three types of kurtosis

coefficient of kurtosis less than 3 are *platykurtic* – flatter in shape than the standard normal distribution. Each of these types of coefficients of kurtosis is illustrated in Fig. 9.11. Sometimes, the sample *coefficient of kurtosis* (SCK) can be redefined as

$$\text{SCK}' = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{\left[\Sigma (X_i - \bar{X})^2\right]^2} - 3 \tag{9.26}$$

The value for the redefined CK for a normal distribution is 0 instead of 3.

If X is lognormally distributed, then from Sect. 7.6, the mean and variance of X can be defined as

$$\mu'_1 = \mu_X = e^{\mu + 1/2\sigma^2} \tag{7.6}$$

$$\mu_2 = \sigma_X^2 = e^{2\mu + \sigma^2} \left(e^{\sigma^2} - 1\right) \tag{7.7}$$

where $\mu = E(\log X)$ and $\sigma^2 = \text{Var}(\log X)$.

From Eqs. 7.6 and 7.7, the coefficient of variation (η) for X can be defined as

$$\eta = \left(e^{\sigma^2} - 1\right)^{1/2} \tag{9.27}$$

The third and fourth moments about the mean for lognormal distributions are

$$\mu_3(\text{skewness of } X) = (\mu_X)^3 \left(\eta^6 + 3\eta^4\right) \tag{9.28}$$

$$\mu_3(\text{kurtosis of } X) = (\mu_X)^4\left(\eta^{12} + 6\eta^{10} + 15\eta^8 + 16\eta^6 + 3\eta^4\right) \tag{9.29}$$

where $\eta^2 = e^{\sigma^2} - 1$. (See Appendix 4 for the derivation of Eqs. 9.28 and 9.29.)

Substituting μ_1, μ_3, and μ_4 into Eqs. 4.16 and 9.24, we obtain the following equations for the coefficient of skewness (CS) and the coefficient of kurtosis (CK):

$$\text{CS} = \eta^3 + 3\eta \tag{9.30}$$

$$\text{CK} = \eta^8 + 6\eta^6 + 15\eta^4 + 16\eta^6 \tag{9.31}$$

where $\eta^2 = e^{\sigma^2} - 1$.

From Eqs. 9.28, 9.29, 9.30, and 9.31, we know that the coefficient of variation is the key variable in determining the magnitude of both skewness and kurtosis for a lognormal distribution.

In the next section, we will see how Eqs. 4.1, 4.7, 4.12, 4.16a, and 9.26 are applied with data on stock rates of return.

9.8 Analyzing the First Four Moments of Rates of Return of the 30 DJI Firms

In Table 9.1, we have listed the first four moments of the monthly returns of the 30 companies included in the Dow Jones Industrial (DJI) Average. These moments describe the central tendency, variability, asymmetry, and peakedness of the monthly return distributions between January 1990 and December 2009, inclusive. The mean column gives us a measure of central tendency. The average mean of these 30 companies is .0013. The highest monthly return mean was from McDonald's, followed by Disney, Verizon Inc., and United Technologies Corp. The lowest performances were for Bank of America and Alcoa, which had returns of −.037 and −.02, respectively.

The measure of variability is given by the standard deviation. The average standard deviation was .0813. The two companies that showed the highest variability were Bank of America and Alcoa. The lowest variability was achieved by Johnson & Johnson, followed by McDonald's.

In fact, we usually observe that higher rates of return are associated with higher levels of risk. Note that these companies that generated high rates of return tend to have high variability. The principle is simple: the higher the return you seek, the more risk you have to take. There is a trade-off between risk and return, which will be discussed in Chap. 21 in some detail.

The skewness can be used to evaluate the stock's upside potential and downside risk. Positive skewness indicates the upside potential for a stock, because such a stock has a greater probability of very large payoffs. On the other hand, negative

Table 9.1 Statistical estimates for the Dow Jones 30 industrial firms (January 1990–December 2009)

		Monthly statistical estimates					
	Company name	*n*	*Mean*	*Standard deviation*	*Skewness*	*Kurtosis*	*Coefficient of variation*
1	3M Co.	120	0.0019	0.0665	−0.4896	0.549	35.762
2	Alcoa Inc.	120	−0.0211	0.1765	−3.0010	13.522	−8.373
3	American Express	120	−0.0094	0.1230	0.1616	4.211	−13.032
4	AT&T	120	0.0055	0.0616	−1.0458	0.834	11.191
5	Bank of America	120	−0.0370	0.2225	−2.5486	10.608	−6.021
6	Boeing	120	−0.0036	0.0965	−1.0373	1.596	−26.544
7	Caterpillar Inc	120	0.0029	0.1279	−1.9207	6.756	43.496
8	Chevron	120	0.0087	0.0624	−0.7818	0.619	7.168
9	Cisco	120	−0.0010	0.0892	−0.6297	0.612	−86.495
10	Coca-Cola	120	0.0093	0.0510	−0.9161	3.869	5.483
11	E.I. du Pont de Nemours	120	0.0024	0.0872	−0.7714	1.938	36.005
12	Exxon	120	0.0048	0.0520	−0.4104	−0.067	10.810
13	General Electric	120	−0.0135	0.1126	−1.0687	2.134	−8.322
14	Hewlett-Packard	120	0.0044	0.0752	−1.0900	1.269	17.139
15	Home Depot	120	−0.0028	0.0768	−0.3816	0.091	−27.759
16	Intel	120	−0.0041	0.0855	−0.9769	0.878	−20.822
17	IBM	120	0.0093	0.0603	−1.6744	5.742	6.494
18	Johnson & Johnson	120	0.0019	0.0443	−0.9569	1.848	22.763
19	JPMorgan and Chase	120	−0.0016	0.0991	−0.6226	0.947	−61.561
20	Kraft Foods	120	0.0030	0.0623	−1.2024	2.413	20.474
21	McDonald's	120	0.0173	0.0461	−0.2755	−0.159	2.667
22	Merck	120	0.0086	0.0778	−0.2400	0.473	9.049
23	Microsoft	120	0.0058	0.0802	0.0533	0.482	13.939
24	Pfizer	120	0.0012	0.0638	−0.1491	0.028	51.410
25	Procter and Gamble	120	0.0051	0.0489	−0.3265	0.172	9.660
26	Traveler's Companies Inc.	120	0.0072	0.0537	−0.1024	1.817	7.451
27	United Technologies Group	120	0.0094	0.0616	−0.4312	−0.035	6.564
28	Verizon	120	0.0095	0.0571	−0.0976	−0.544	5.986
29	Walmart	120	0.0051	0.0473	−0.3584	1.539	9.333
30	Walt Disney	120	0.0109	0.0700	−0.3271	0.961	6.431
	Mean		0.0013	0.0813	−0.7873	2.170	2.678

skewness is associated with downside risk; it indicates that the stock has a greater probability of very small payoffs.[9] There are 28 companies in Table 9.1 exhibit the downside risk associated with negative skewness. The others, American Express and Microsoft, exhibit the upside potential associated with positive skewness.

[9] This is so because a positively skewed distribution has more observations above the mode and a negatively skewed distribution more observations below.

The kurtosis column shows that 26 companies here have a leptokurtic distribution (kurtosis ratio > 0)[10]; these companies have more monthly returns concentrated near the mean. Only four companies have a distribution close to platykurtic (kurtosis ratio < 0): Exxon with SCK$' = -.067$, McDonald's with SCK$' = -.159$, United Technologies Corp with SCK$' = -.035$, and Verizon with SCK$' = -.544$.

The last column, showing the coefficient of variation, enables us to compare monthly returns for the different companies. Remember that the coefficient of variation is a unitless figure that expresses the standard deviation as a percentage of the mean. High coefficients of variation show volatile monthly returns. The companies that show high volatility are Pfizer, 3M, Caterpillar, and E.I du Pont de Nemours. The companies with the lowest volatility are Cisco, JPMorgan Chase, Home Depot, and Boeing.

9.9 Summary

In this chapter, we discussed five continuous distributions. Four of these – Student's t distribution and the exponential, F, and χ^2 distributions – are closely related to the normal distribution discussed in Chap. 7. These five distributions, along with the normal and lognormal distributions, are the primary distributions we will use throughout the rest of the text for conducting statistical analyses such as determination of confidence intervals, hypothesis testing, and goodness-of-fit tests.

In Chaps. 11, 12, 13, 14, and 15, we will begin to apply these distributions in alternative statistical analyses.

Questions and Problems

1. Briefly discuss the cumulative distribution function of the uniform distribution presented in Fig. 9.2.
2. Briefly discuss the relationship between the Poisson distribution and the exponential distribution.
3. X is normally distributed, and the sample variance $s^2 = 20$ is calculated from 20 observations. Calculate $E(s^2)$ and $\text{Var}(s^2)$.
4. W is a normally distributed random variable with mean 0 and variance 1, and V is a χ^2-distributed random variable with degrees of freedom $(n - 1)$. How can both t and F distributions be defined in terms of the variables W and V?
5. Briefly discuss how F statistics can be used to test the difference between two sample variances.
6. Briefly discuss how mean, variance, skewness, kurtosis, and the coefficient of variation can be used to analyze stock rates of return.

[10] We use Eq. 9.26 to calculate the coefficient of kurtosis.

7. Suppose a random variable X can take on only values in the range from 2 to 10 and that the probability that the variable will assume any value within any interval in this range is the same as the probability that X will assume another value in another interval of similar width in the range. What is the distribution of X? Draw the probability density function for X.
8. Use the information given in question 7 to find $P(3 \leq X \leq 7)$.
9. Use the information given in question 7 to find $P(X \leq 8)$.
10. Use the information given in question 7 to find $P(X < 2 \text{ or } X > 10)$.
11. Draw the cumulative distribution function for the distribution given in question 7.
12. Suppose a random variable X is best described by a uniform distribution with $a = 8$ and $b = 20$.

    (a) Find $f(x)$.
    (b) Find $F(x)$.
    (c) Find the mean and variance of X.

13. Suppose a random variable Y is best described by a uniform distribution with $a = 3$ and $b = 32$.

    (a) Find $f(y)$.
    (b) Find $F(y)$.
    (c) Find the mean and variance of Y.

14. A very observant art thief (who should probably be teaching statistics instead) notices that the frequency of security guards passing by a museum is uniformly distributed between 15 and 60 min. Therefore, if X denotes the time (in minutes) before the guard passes by, the probability density function of X is

$$f_x(x) = \begin{cases} 1/(60-15) & \text{for } 15<x<60 \\ 0 & \text{for all other values of } x \end{cases}$$

    (a) Draw the probability density function.
    (b) Find and draw the cumulative distribution function.

15. Use the information given in question 14.

    (a) Find the probability that the guard passes by within 35 min of the thief's arrival.
    (b) Find the probability that the guard does not pass by within 30 min.
    (c) Find the probability that the guard passes by between 30 and 45 min after the thief's arrival.

16. An art dealer at an auction believes that the bid on a certain painting will be a uniformly distributed random variable between \$500 and \$2,000.

    (a) What is the probability density function for this random variable?
    (b) Find the probability that the painting will sell for less than \$675.
    (c) Find the probability that the painting will sell for more than \$1,000.

17. Suppose X has an exponential distribution with $\lambda = 5$. Find the following probabilities:

    (a) $P(X > 4)$
    (b) $P(X > .7)$
    (c) $P(X > .50)$

18. Suppose X has an exponential distribution with $\lambda = 4$. Find the following probabilities:

    (a) $P(X \leq .3)$
    (b) $P(X \leq .5)$
    (c) $P(X \leq 1.6)$

19. Suppose X has an exponential distribution with $\lambda = \frac{1}{3}$. Find the following probabilities:

    (a) $P(3 \leq X \leq 5)$
    (b) $P(5 \leq X \leq 10)$
    (c) $P(2 \leq X \leq 1)$

20. Suppose the random variable X is best approximated by an exponential distribution with $\lambda = 8$. Find the mean and the variance of X.
21. Suppose the random variable Y is best approximated by an exponential distribution with $\lambda = 3$. Find the mean and the variance of Y.
22. Briefly compare the normal distribution discussed in Chap. 7 with the t distribution discussed in this chapter.
23. Find t_α for the following:

    (a) $\alpha = .05$ and $v = 10$
    (b) $\alpha = .025$ and $v = 4$
    (c) $\alpha = .10$ and $v = 7$

24. Find the value t_0 such that

    (a) $P(t \geq t_0) = .025$, where $v = 6$
    (b) $P(t \geq t_0) = .05$, where $v = 12$
    (c) $P(t \leq t_0) = .10$, where $v = 9$

25. Find the value t_0 such that

    (a) $P(t \leq t_0) = .10$, where $v = 25$
    (b) $P(t \geq t_0) = .025$, where $v = 14$
    (c) $P(t \leq t_0) = .01$, where $v = 17$

26. Find the following probabilities for the t distributions.

    (a) $P(t > 3.078)$ if $v = 1$
    (b) $P(t < 1.943)$ if $v = 6$
    (c) $P(t > 2.492)$ if $v = 24$

27. Find the following probabilities for the t distributions.

    (a) $P(t > 1.734)$ if $v = 18$
    (b) $P(t > 1.943)$ if $v = 6$
    (c) $P(t < 1.645)$ if $v = \infty$

28. Find the following $\chi^2_{\alpha,v}$ values.

    (a) $\alpha = .05$ and $v = 25$
    (b) $\alpha = .025$ and $v = 5$
    (c) $\alpha = .10$ and $v = 50$
    (d) $\alpha = .01$ and $v = 60$

29. Find the following $\chi^2_{\alpha,v}$ values.

    (a) $\alpha = .025$ and $v = 30$
    (b) $\alpha = .0l$ and $v = 70$
    (c) $\alpha = .10$ and $v = 10$
    (d) $\alpha = .01$ and $v = 20$

30. Find the following probabilities.

    (a) $P(\chi^2 > 10.8564)$ when $v = 24$
    (b) $P(\chi^2 < 10.8564)$ when $v = 24$
    (c) $P(\chi^2 < 48.7576)$ when $v = 70$
    (d) $P(\chi^2 > 59.1963)$ when $v = 90$

31. Find the following probabilities.

    (a) $P(\chi^2 \leq 3.84146)$ when $v = 1$
    (b) $P(\chi^2 \geq 15.9871)$ when $v = 10$
    (c) $P(\chi^2 < 140.169)$ when $v = 100$
    (d) $P(\chi^2 > 1.61031)$ when $v = 5$

32. Find the following $F_{v_1,v_2,\alpha}$ values.

    (a) $v_1 = 8$, $v_2 = 10$, and $\alpha = .01$
    (b) $v_1 = 3$, $v_2 = 11$, and $\alpha = .005$
    (c) $v_1 = 12$, $v_2 = 9$, and $\alpha = .05$
    (d) $v_1 = 24$, $v_2 = 19$, and $\alpha = .025$

33. Find the following $F_{v_1,v_2,\alpha}$ values.

    (a) $v_1 = 10$, $v_2 = 10$, and $\alpha = .05$
    (b) $v_1 = 15$, $v_2 = 3$, and $\alpha = .01$
    (c) $v_1 = 12$, $v_2 = 15$, and $\alpha = .025$
    (d) $v_1 = 20$, $v_2 = 10$, and $\alpha = .005$

34. Find the probabilities, given v_1 and v_2 as shown.

    (a) $v_1 = 1$ and $v_2 = 3$; $P(F > 17.44)$
    (b) $v_1 = 3$ and $v_2 = 1$; $P(F > 864.2)$

(c) $v_1 = 3$ and $v_2 = 1$; $P(F < 215.7)$
(d) $v_1 = 30$ and $v_2 = 12$; $P(F < 4.33)$

35. Using the MINITAB program, find the probabilities, given v_1 and v_2 as shown.

(a) $v_1 = 120$ and $v_2 = 120$; $P(F > 1.35)$
(b) $v_1 = 00$ and $v_2 = \infty$; $P(F > 1.00)$
(c) $v_1 = 6$ and $v_2 = 17$; $P(F < 3.28)$
(d) $v_1 = 3$ and $v_2 = 23$; $P(F > 4.76)$

36. Find the probability that an exponentially distributed random variable X with mean $1/\lambda = 8$ will take on the values:

(a) Between 2 and 7
(b) Less than 9
(c) Greater than 6
(d) Between 1 and 15

37. Suppose the lifetime of a television picture tube is distributed exponentially with a standard deviation of 1,400 h. Find the probability that the tube will last:

(a) More than 3,000 h
(b) Less than 1,000 h
(c) Between 1,000 and 2,000 h

38. Suppose the time you wait at a bank is exponentially distributed with mean $1/\lambda = 12$ min. What is the probability that you will wait between 10 and 20 min?
39. Suppose the length of time people wait at a fast-food restaurant is distributed exponentially with a mean of 1/7 min. Use MINITAB to answer the following questions.

(a) What percentage of people will be served within 4 min?
(b) What percentage of people will be served between 3 and 8 min after they arrive?
(c) What percentage of people will wait more than 9 min?

40. Suppose the length of time a student waits to register for courses is distributed exponentially with a mean of 1/15 min.

(a) What percentage of students will register within 10 min?
(b) What percentage of students will register after waiting between 10 and 20 min?
(c) What percentage of students will wait more than 20 min to register?

41. Suppose a random variable is distributed as an x^2 distribution with n degrees of freedom. Consider the probability $P(x^2 \leq 9)$. Explain the relationship between the probability and the degrees of freedom.
42. Suppose a random variable is distributed as Student's t distribution with $(n - 1)$ degrees of freedom. Consider the probability $P(t \geq .7)$. Explain the relationship between the probability and the degrees of freedom.

43. The incomes of families in a town are assumed to be uniformly distributed between $15,000 and $85,000. What is the probability that a randomly selected family will have an income above $40,000?
44. At an antiques auction, the winning bids were found to be uniformly distributed between $500 and $2,500. What is the probability that a winning bid was less than $1,000? What is the probability that a winning bid was between $750 and $1,500?
45. The manager of a department store notices that the amount of time a customer must wait before being helped is distributed uniformly between 1 and 4 min. Find the mean and variance of the time a customer must wait to be helped.
46. A quality control expert for the Healthy Time Cereal Company notices that in a 16-oz package of cereal, the amount in the box is uniformly distributed between 15.3 and 17.1 oz. Find the mean and standard deviation for the weight of this cereal in a package of cereal.
47. The shelf life of hearing aid batteries is found to be approximated by an exponential distribution with a mean of 1/12 day. What fraction of the batteries would be expected to have a shelf life greater than 9 days?
48. A computer programmer has decided to use the exponential distribution to evaluate the reliability of a computer program. After 10 programming errors were found, the time (measured in days) to find the next error was determined to be exponentially distributed with a $\lambda = .25$.

    (a) Graph this distribution.
    (b) Find the mean time required to find the 11th error.

49. Use the information given in question 48 to find the probability that it will take more than 5 days to find the 11th error. Find the probability that it will take between 3 and 10 days to find the 11th error.
50. An advertising executive believes that the length of time a television viewer can recall a commercial is distributed exponentially with a mean of .25 days. Find how long it will take for 75 % of the viewing audience to forget the commercial.
51. Use the information given in question 50 to find the proportion of viewers who will be able to recall the commercial after 7 days.
52. An investment advisor believes that the rate of return for Horizon Company's stock is uniformly distributed between 3 % and 12 %. Find the probability that the return will be greater than 5 %. Find the probability that the return will be between 6 % and 8 %.
53. The mean life of a computer's hard disk is found to be exponentially distributed with a mean of 12,000 h. Find the proportion of hard disks that will have a life greater than 20,000 h.
54. Suppose the life of a car battery is assumed to be uniformly distributed between 3.9 and 7.3 years. Find the mean and variance of the life of a car battery.
55. Use the information given in question 54 to find the probability that the life of the car battery will be greater than 5 years. Find the probability that the life of the battery will be between 4 and 6 years.

56. The chief financial officer at Venture Corporation believes that an investment in a new project will have a cash flow in year one that is uniformly distributed between $1 million and $10 million. What is the probability that the cash flow in year one will be greater than $1.7 million?
57. A hospital collects data on the number of emergency room patients in during a certain period. It is estimated that in an hour, the average number of emergency room patients to arrive is 1.2. If the time between two consecutive arrivals of patients follows an exponential distribution, what is the probability that a patient will show up in the next hour?
58. The campus bus at Haverford College is scheduled to arrive at the business school at 8:00 a.m. Usually, the bus arrives at the bus stop during the interval 7:56–8:03. Assume that the arrival time follows a uniform distribution.

    (a) What is the probability that the bus arrives at the business school before 8:00?
    (b) What is the average arrival time?
    (c) What is the standard deviation of arrival time?

59. A gas station's owner found that about two cars come into the station every minute. If the arrival time follows an exponential distribution, what is the probability that the next car will arrive in 1.5 min?
60. A college professor gives a standardized test to her students every semester. She finds that the students' grades follow a uniform distribution with 100 points as the maximum and 65 points as the minimum.

    (a) Find the mean score.
    (b) Compute the standard deviation of the score.
    (c) If the passing grade is 70, what percentage of students will fail the course?

61. Suppose the weight of a football team is uniformly distributed with a minimum weight of 175 lb and a maximum weight of 285 lb.

    (a) Find the mean weight of the team.
    (b) Compute the standard deviation of the weight.
    (c) Find the percentage of players with a weight of less than 195 lb.

62. Briefly explain how the mean, standard deviation, coefficient of variation, and skewness can be used to analyze the returns of IBM and Boeing in Table 9.1.
63. A bank manager finds that about six customers enter the bank every 5 min. If the customer arrival time follows an exponential distribution, what is the probability that the next customer will arrive in 2 min?
64. Suppose the life of a steel-belted radial tire is uniformly distributed between 30,000 and 45,000 miles.

    (a) Find the mean tire life.
    (b) Find the standard deviation of tire life.
    (c) What percentage of these tires will have a life of more than 40,000 miles?

65. Briefly discuss the relationship among t, χ^2, and F distributions.

66. Given $v_1 = 5$ and $\alpha = .05$, find v_2 for the following F values.

    (a) 5.05
    (b) 3.33
    (c) 2.53

67. In their study, Vardeman and Ray (*Technometrics*, May 1985, pp. 145–150) found that the number of accidents per hour at an industrial plant is exponentially distributed with a mean $\lambda = .5$. Use the formula $f(t) = \lambda e^{-\lambda t}$ to determine each of the following.

    (a) $f(1)$
    (b) $f(4)$
    (c) $E(t)$

68. Suppose there is a sample of 30 items drawn from a normal population. Find the probability that the sample variance exceeds 36.6869.

69. Suppose there are two independent normal populations with population variances $\sigma^2_1 = 4.5$ and $\sigma^2_2 = 2.5$, respectively. Two random samples of sizes s^2_1 and s^2_2, respectively, are drawn from the two normal populations with sample variances s^2_1 and s^2_2, respectively.

    (a) What is the probability that the ratio s^2_1/s^2_2 is greater than 4.230?
    (b) What is the probability that the ratio s^2_1/s^2_2 is greater than 6.066?
    (c) What is the probability that the ratio s^2_1/s^2_2 is less than 0.5263?

70. A random sample of size 7 is drawn from a population with population variance $\sigma^2 = 2.5$.

    (a) Determine the probability that the variance of the sample is greater than 7.008.
    (b) Determine the probability that the population mean is less than 0.3634.

71. The following random sample is taken from a normal population.

94	72	43	69	28	63	93	54	77	58

    (a) If the population man is $\mu = 60$, what is t statistics for the sample?
    (b) If the population man is $\mu = 55$, what is t statistics for the sample?
    (c) What is the degree of freedom of the t statistics in (a)?

Project II: Project for Probability and Important Distributions

1. Use rates of return data presented in Table 2.4 to do the following:

    (a) Use either MINITAB or Microsoft Excel to calculate:

        1. Mean
        2. Standard deviation

(continued)

Project II: (continued)

3. Coefficient of variation
4. Skewness
5. Kurtosis

(b) Analyze the statistical results of (a).
(c) Use both the standard deviation for JNJ and Merck calculated in (a) and the following information to calculate the call option and put option values for JNJ and Merck:

$$S = \$50 \quad X = 45 \quad r = 6\% \quad T = .6$$

2. Use MINITAB and the statistical estimates for JNJ and Merck obtained in (a) to calculate the mean and the variance of a portfolio with the following weights:
   1. $w_1 = .4$ and $w_2 = .6$
   2. $w_1 = .2$ and $w_2 = .8$
   3. $w_1 = .3$ and $w_2 = .7$
   4. $w_1 = .1$ and $w_2 = .9$
3. Download monthly adjusted close price data of JNJ from Yahoo Finance during the period from January 2005 to current month:
   (a) Calculate monthly rates of return of JNJ.
   (b) Redo 1a–c.

Appendix 1: Derivation of the Mean and Variance for a Uniform Distribution

On the basis of the definitions of $E(X)$ and $E(X^2)$ for a continuous variable given in Appendix 1 of Chap. 7, we can derive the mean and the variance of a uniform distribution as follows. First, substituting Eq. 9.1 into Eq. 7.22, we get

$$\begin{aligned} E(X) &= \int_a^b xf(x)dx = \int_a^b \frac{x}{b-a}dx \\ &= \frac{1}{b-a} \cdot \frac{x^2}{2}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2} \end{aligned} \tag{9.32}$$

Then, substituting Eq. 9.1 into Eq. 7.25 yields

$$E(X^2) = \int_a^b x^2 f(x)dx = \frac{1}{b-a}\int_a^b x^2 dx = \frac{1}{b-a}\cdot\frac{x^3}{3}\Big|_a^b$$
$$= \frac{b^3-a^3}{3(b-a)} = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3} \tag{9.33}$$

Finally, substituting Eqs. 9.32 and 9.33 into the definition of variance given in Eq. 7.24, we obtain

$$\sigma_X^2 = E(X^2) - [E(X)]^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2$$
$$= \frac{4b^2+4ab+4a^2-3b^2-6ab-3a^2}{12} = \frac{(b-a)^2}{12} \tag{9.34}$$

This implies that $\sigma_X = (b-a)/\sqrt{12}$.

The following example shows how the formulas for both the mean and the variance of a continuous variable, as discussed in Appendix 1 of Chap. 7, can be applied for a uniform distribution.

Example 9.4 Calculating the Mean and Variance of a Uniform Distribution. Let us look at an example of a continuous random variable in terms of the uniform distribution. Consider the density function of Eq. 9.35 as depicted in Fig. 9.12:

$$f(x) = \begin{cases} 1.55 - .06x & \text{if } 20 \le x \le 25 \\ 0 & \text{otherwise} \end{cases} \tag{9.35}$$

For every value of x between 20 and 25, we get $f(x) > 0$, and for every x value outside of this range, we have $f(x) = 0$. Therefore, for every x, we have $f(x) \ge 0$. Furthermore, the area under the curve equals 1:

$$\int_{20}^{25}(1.55-.06x)dx = 1.55x\Big|_{20}^{25} - \frac{.06x^2}{2}\Big|_{20}^{25} = 1$$

This confirms that $f(x)$ is a density function. Now, let us calculate the expected value and variance of X:

$$E(X) = \int_{20}^{25} xf(x)dx = \int_{20}^{25} x(1.55-.06x)dx$$
$$= \frac{1.55x^2}{2}\Big|_{20}^{25} - \frac{.06x^2}{3}\Big|_{20}^{25}$$
$$= 174.375 - 152.5 = 21.875$$

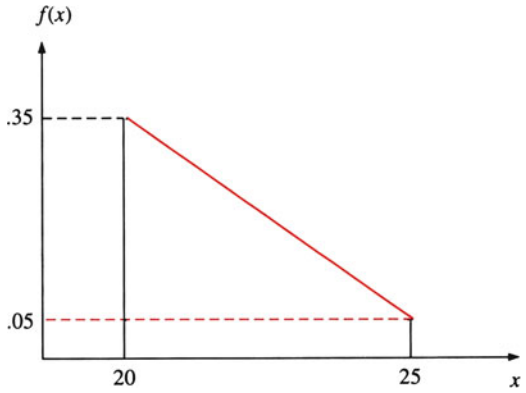


Fig. 9.12 The density function $f(x)$

Next, let us calculate $E(X^2)$:

$$\begin{aligned} E(X^2) &= \int_{20}^{25} x^2(1.55 - .06x)dx \\ &= \frac{1.55x^3}{3}\Big|_{20}^{25} - \frac{.06x^4}{4}\Big|_{20}^{25} \\ &= 3939.583 - 3459.375 = 480.208 \end{aligned}$$

From this result, we obtain

$$V(X) = E(X^2) - (EX)^2 = 480.208 - (21.875)^2 = 1.692$$

To find the probability, such as $P(22 \leq X \leq 24.5)$, we calculate

$$\begin{aligned} P(22 \leq X \leq 24.5) &= \int_{22}^{24.5} f(x)dx = \int_{22}^{24.5} (1.55 - .06x) \\ &= 1.55x\Big|_{22}^{24.5} - \frac{.06x^2}{2}\Big|_{22}^{24.5} \\ &= 3.875 - 3.4875 = .3875 \end{aligned}$$

Appendix 2: Derivation of the Exponential Density Function

The cumulative distribution function (CDF) for the first event to occur in time interval t can be written as

$$\begin{aligned} P(T \leq t) &= P(\text{wait until next arrival} \leq t) \\ &= P(\text{at least one arrival in time } t) \\ &= 1 - P(\text{non-arrival in time } t) \end{aligned} \tag{9.36}$$

where T is the random variable of which t is a specific value. P(non-arrival in time t) can be obtained by letting $x = 0$ in the Poisson function as defined in Eq. 6.16. We obtain P(non-arrival in time interval [0, t]) as

$$\begin{aligned} f(0) = P(T \geq t) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} &= e^{-\lambda t} \text{ for } t \geq 0 \\ &= 0 \quad \text{for } t < 0 \end{aligned} \tag{9.37}$$

where λ denotes the mean rate at which events occur over time. Substituting Eq. 9.37 into Eq. 9.36, we obtain the CDF as

$$F(t) = P(T \leq t) = 1 - e^{-\lambda t} \tag{9.38}$$

If we differentiate $F(t)$ with respect to t, we obtain the PDF as[11]

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t}, \quad t \geq 0 \\ &= 0, \qquad t < 0 \end{aligned} \tag{9.39}$$

The probability that the waiting time lies between a and b is

$$P(a)) = \int_a^b \lambda e^{-\lambda t} \tag{9.40}$$

From the definition of $E(t)$ in Appendix 1 of Chap. 7, we obtain

$$E(T) = \int_{-\infty}^{\infty} tf(t)dt = \lambda \int_0^{\infty} te^{-\lambda t} dt$$

The integral can be evaluated by parts. Let $U = t$ and $dv = e^{-\lambda t}\, dt$, so $dU = dt$ and $v = -e^{-\lambda t}/\lambda$. Then

$$\begin{aligned} E(T) &= \lambda \left[\left(-te^{-\lambda t}/\lambda \right)_0^{\infty} + \frac{1}{\lambda} \int_0^{\infty} e^{-\lambda t} dt \right] \\ &= \lambda \left[(-0 + 0) + \frac{1}{\lambda^2} (-0 + 1) \right] = \frac{1}{\lambda} \end{aligned}$$

[11] This is because

$$\frac{dF(t)}{dt} = \frac{d\left(1 - e^{-\lambda t}\right)}{dt} = 0 - \left[\frac{d(-\lambda t)}{dt} \right] e^{-\lambda t} = \lambda e^{-\lambda t}$$

Similarly, we can prove that

$$\text{Var}(T) = \frac{1}{\lambda^2} \tag{9.41}$$

This appendix shows how a mean value formula of a continuous variable, which was discussed in Appendix 1 of Chap. 7, can be applied to an exponential distribution.

Example 9.5 The Average Time Required to Find the Next Computer Program Error. In finding and correcting errors in a computer program (debugging) and determining the program's reliability, Schick and others have noted the importance of the distribution of the time until the next program error is found. The cumulative exponential probability function of Eq. 9.37 is most useful in analyzing this problem.

By using the computer debugging data supplied by the US Navy, Schick (1974, *Decision Sciences*, Vol. 5, pp. 529–544) estimated the value of λ. After 26 of 31 program errors were found, Schick estimated λ to be .042. Accordingly, $1/\lambda = 23.8$ days. This means that the average time it would take to find 1 of the remaining errors (the 27th error) would be about 24 days. From this information, we can estimate, for example, that the probability of taking 50 or more days to find the next error is

$$P(T \geq 50) = e^{-(.042)(50)} = e^{-2.1} = .1125.$$

The second equality is obtained by using Table A7 in Appendix A.

Example 9.6 The Probability of Truck Arrivals. Rutgers Trucking Company had 15,600 trucks to unload at the receiving warehouse during the last calendar year. The warehouse was open from 8 a.m. to 8 p.m. each weekday. There was no noticeable pattern of truck arrivals each day. It is known that approximately five trucks arrived to unload cargo each hour. What is the probability that on September 20, 1991, the first truck arrived between 8:15 and 8:30 a.m.?

To use exponential distribution to solve this problem, we first use a time interval of 15 min (8:15–8:30) for which $\lambda = (5/60)(15) = 1.25$.

Substituting $\lambda = 1.25$, $a = 1$, and $b = 2$ into Eq. 9.40,[12] we obtain the probability that the first truck arrived between 8:15 and 8:30 a.m.:

$$\begin{aligned} P(1<T<2) &= \int_1^2 e^{-1.25t}(1.25\text{dt}) = -e^{-1.25t}\Big|_1^2 \\ &= -e^{-2.5} + e^{-1.25} = .2 \end{aligned}$$

[12] We regard 15 min as 1 time unit that can be expressed as a time interval between $a = 1$ and $b = 2$.

Appendix 3: The Relationship Between the Moment About the Origin and the Moment About the Mean

Let $k = 1$ in Eq. 9.22. Then

$$\mu_1 = E(X - \mu'_1) = E(X) - \mu'_1 = 0$$

This implies that the first moment about the population mean is zero.

Alternatively, if we let $k = 2$ in Eq. 9.22 and let $\mu_1 = \mu_1$, we obtain

$$\begin{aligned}\mu_2 &= E(X - \mu'_1)^2 = E\left(X^2 - 2X\mu'_1 + \mu'_1\right) \\ &= E\left(X^2\right) - 2\mu'_1 E(X) + \mu'^2_1 = \mu'_2 - \mu'^2_1\end{aligned} \tag{9.42}$$

where μ'_2 and μ'_1 are second and first moments, respectively. Equation 9.42 is identical to Eq. 7.24 in Appendix 1 of Chap. 7. It is a shortcut formula to calculate variance.

Now, if we *let* $k = 3$ in Eq. 9.23 and substitute $\mu_1 = \mu'_1$, we obtain

$$\begin{aligned}\mu_3 =& E\left(X - \mu'_1\right)^3 = E\left(X^3 - 3X^2\mu'_1 + 3X\mu'^2_1 - \mu'^3_1\right) \\ =& E(X^3) - 3\mu'_1 E(X^2) + 3\mu'^2_1 E(X) - \mu'^3_1 \\ =& \mu'_3 - 3\mu'_1\mu'_2 + 2\mu'^3_1\end{aligned} \tag{9.43}$$

where μ'_1 and μ'_2 are defined in Eq. 9.42 and μ'_3 is the third moment about the origin.

Finally, letting $k = 4$ in Eq. 9.22 and substituting $\mu_1 = \mu_1$, we obtain

$$\begin{aligned}\mu_4 &= E\left(X - \mu'_1\right)^4 \\ &= E\left(X^4 - 4X^3\mu'_1 + 6X^2\mu'^2_1 - 4E(X)\mu'^3_1 + \mu'^4_1\right) \\ &= E\left(X^4\right) - 4E\left(X^3\right)\mu'_1 + 6E\left(X^2\right)\mu'^2_1 - 4E(X)\mu'^3_1 + \mu'^4_1 \\ &= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1\end{aligned} \tag{9.44}$$

where μ'_1, μ'_2, and μ'_3 have been defined in Eq. 9.43 and μ'_4 is the fourth moment about the origin.

In Appendix 4, Eqs. 9.42, 9.43, and 9.44 will be used to derive variance, skewness, and kurtosis of the lognormal distribution.

Appendix 4: Derivations of Mean, Variance, Skewness, and Kurtosis for the Lognormal Distribution

Following Aitchison and Brown (1963), we express the moments about the origin for the lognormal distribution as

$$\mu'_k = e^{k\mu+1/2k^2\sigma^2}, \quad k = 1,\ 2, \cdots \tag{9.45}$$

In accordance with definitions given in Appendix 3, the mean, variance skewness, and kurtosis of a lognormal distribution can be derived as follows.

Mean

Substituting $k = 1$ into Eq. 9.45 yields

$$\mu'_1 = e^{\mu+1/2\sigma^2}$$

This is Eq. 7.6.

Variance

Substituting $\mu'_2 = e^{2\mu+2\sigma^2}$ and $\mu'_1 = e^{\mu+1/2\sigma^2}$ into Eq. 9.42 in Appendix 3, we obtain

$$e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

This is Eq. 7.7.

Skewness

Substituting μ'_1, μ'_2, and $\mu'_3 = e^{3\mu+9/2\sigma^2}$ into Eq. 9.43 gives

$$\begin{aligned}\mu_3 =& (\mu'_1)^3\left[e^{3\sigma^2} - 3e^{\sigma^2} + 2\right]\\ =& (\mu'_1)^3\left[\left(e^{3\sigma^2} - 3e^{2\sigma^2} + 3e^{\sigma^2} - 1\right) + 3\left(e^{2\sigma^2} - 2e^{\sigma^2} + 1\right)\right]\\ =& (\mu'_1)^3\left(\eta^6 + 3\eta^4\right)\end{aligned}$$

where $\eta^2 = e^{\sigma 2} - 1$. This is Eq. 9.28.

Kurtosis

Substituting μ'_1, μ'_2, μ'_3, and $\mu'_4 = e^{3\mu+8\sigma^2}$ into Eq. 9.44, we get

$$\mu_4 = (\mu'_1)^4 \left[e^{6\sigma^2} - 4e^{3\sigma^2} + 6e^{\sigma^2} - 3\right]$$

By considerable mathematical rearrangement of terms, it can be shown that

$$\mu_4 = (\mu'_1)^4 \left[\eta^{12} + 6\eta^{10} + 15\eta^8 + 16\eta^6 + 3\eta^4\right]$$

where $\eta^2 = e^{\sigma^2} - 1$. This is Eq. 9.29.

Appendix 5: Noncentral χ^2 and the Option Pricing Model

From Eq. 9.6, we know that $Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma_x}\right)^2$ is distributed as χ^2 with n degree of freedom. This is a central χ^2 distribution. It can be shown that $Y' = \sum_{i=1}^{n} X_i^2$ is distributed as noncentral χ^2 with n degree of freedom and a noncentral parameter[13]

$$\lambda = \frac{1}{2}\sum \mu_i^2.$$

If $\mu = 0$, the distribution of Y' reduces to the central χ^2 distribution.

The option pricing model defined in Appendix 2 of Chap. 7 assumed that the variance of stock rate of return (σ^2) is constant. If the variance of stock rate of return is a function of stock price per share, $\sigma^2 S^{\beta-2}$, then the option pricing model defined in Eq. 7.35 can be generalized as[14]

$$C = S\left[1 - \chi^2\left(2n; 2 + \frac{2}{2-\beta}, 2m\right)\right] - Xe^{-r(T-t)}\left[\chi^2\left(2m; \frac{2}{\beta-2}, 2n\right)\right] \quad (\beta<2) \quad (9.46)$$

[13] See Robert V.H. Allen T.C.: Introduction to Mathematical Statistics 4th Edition, pp. 288–290. Macmillan, New York, (1978)

[14] The derivation of this formula can be found in Mark S.: Computing the constant elasticity of variance option pricing formula. J. Finance. **44**, 211–220 (1989)

$$C = S\left[1 - \chi^2\left(2m; \frac{2}{2-\beta}, 2n\right)\right] - Xe^{-r(T-t)}\left[\chi^2\left(2n; 2 + \frac{2}{\beta-2}, 2m\right)\right] \quad ((\beta<2) \quad (9.47)$$

where T = time of expiration of option, t = current time, and r = risk-free rate. $\chi^2(W, V, \lambda)$ is the cumulative noncentral chi-square distribution function with W, V, and, λ being the upper limit of the integral, degree of freedom, and noncentrality, respectively. In addition m, n, and K can be defined as

$$m = KS^{2-\beta}e^{(2-\beta)\mu(T-t)}$$

$$n = KS^{2-\beta}$$

$$K = \frac{2\mu}{\sigma^2(2-\beta)(e^{(2-\beta)\mu(T-t)} - 1)} \quad (9.48)$$

Now, we discuss three possible special cases associated with Eqs. 9.46 and 9.47.

(a) If $\beta = 2$, both m and n approach infinity. Then it can be shown that both Eqs. 9.46 and 9.47 reduce to the well-known Black–Scholes formula as defined in Appendices 2 and 3 of Chap. 7.

(b) If $\beta = 1$, it can be shown that Eqs. 9.46 and 9.47 reduce to

$$C = \left(S - Xe^{-r(T-t)}\right)N(y_1) + \left(S + Xe^{-r(T-t)}\right)N(y_2) + v[n(y_1) - n(y_2)] \quad (9.49)$$

where

$$v = \sigma\sqrt{\frac{1 - e^{-2r(T-t)}}{2r}}$$

$$y_1 = \frac{S - Xe^{-r(T-t)}}{v}$$

$$y_2 = \frac{-S - Xe^{-r(T-t)}}{v}$$

$N(y_1)$ and $N(y_2)$ = cumulative standardized normal distribution function in terms of y_1 and y_2, respectively.
$n(y_1)$ and $n(y_2)$ = standardized normal density function in terms of y_1 and y_2, respectively.

(c) If $\beta = 0$, it can be shown that Eqs. 9.46 and 9.47 reduce to

$$C = SN[q(4)] - Xe^{-r(T-t)}N[q(0)] \quad (9.50)$$

where

$$q(w)=\frac{1+h(h-1)\left(\frac{w+2y}{(w+y)^2}\right)-h(h-1)(2-h(1-3h)\left(\frac{(w+2y)^2}{2(w+y)^4}\right)-\left(\frac{z}{(w+y)}\right)^h}{\left\{2h^2\left(\frac{w+2y}{(w+y)^2}\right)(1-(1-h)(1-3h)\left(\frac{w+2y}{(w+y)^2}\right)\right\}^{\frac{1}{2}}}$$

$$h(w)=1-\frac{2}{3}(w+y)(w+3y)(w+2y)^{-2}$$

$$y=\frac{4rS}{\sigma^2(1-e^{-r(T-t)})} \quad \text{and} \quad z=\frac{4rX}{\sigma^2(e^{-r(T-t)}-1)}$$

The elasticity of variance ($\sigma^2 S^{\beta-2}$) with respect to stock price per share S is

$$\eta_s=\left[\frac{\partial(\sigma^2 S^{\beta-2})}{\partial S}\right]\left[\frac{S}{\sigma^2 S^{\beta-2}}\right]=\left[\frac{(\beta-2)\sigma^2 S^{\beta-2})}{S}\right]\left[\frac{S}{\sigma^2 S^{\beta-2}}\right]=\beta-2 \qquad (9.51)$$

This implies that the option pricing model defined in Eqs. 9.46 and 9.47 is a constant elasticity of variance (CEV) type of OPM.

The CEV type of option pricing model can be reduced to the following special models[15]:

(a) $\beta = 2$, Eqs. 9.46 and 9.47 reduce to the Black–Scholes model.
(b) $\beta = 1$, Eqs. 9.46 and 9.47 reduce to the absolute model as defined in Eq. 9.49.
(c) $\beta = 0$, Eqs. 9.46 and 9.47 reduce to the square root model as defined in Eq. 9.50.

From Appendix 2 of Chap. 6, Appendices 2 and 3 of Chap. 7 and this appendix, we can conclude that the binomial, normal, lognormal, and noncentral χ^2 distributions are basic statistical distributions needed for understanding alternative option pricing models.

[15] See Beckers S.: The constant elasticity of variance model and its implications for option pricing. J. Finance. **35**, 661–673 (1980)

Part III
Statistical Inferences Based on Samples

In the next three chapters, we will discuss statistical inference based on samples and the applications of such statistical inference. So far, our discussion has focused on descriptive statistics, sampling and sampling distributions, and the analytical techniques and distributions used to describe statistical data.

Inferential statistics, on the other hand, is used to make inferences about a population by looking at a subset of that population. In Chapter 10, we continue the discussion of the previous five chapters by looking at point estimation, confidence intervals, and statistical quality control. In Chapter 11, we apply these techniques to testing hypotheses about a population. In Chapter 12, we discuss the analysis of variance for sample data and the use of chi-square tests in analyzing sample data.

Chapter 10
Estimation and Statistical Quality Control

Chapter Outline

Key Terms

Population parameters	Confidence level
Parameter	Probability content
Statistic	Risk probability
Estimate	Significant level
Estimator	Acceptance sampling
Point estimate	Lot
Point estimator	Convenience lots
Point estimation	Single-sampling plans
Unbiasedness	Double-sampling plans
Bias	Control chart
Efficiency	Upper control limit
Consistency	Lower control limit

(continued)

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_10,

Sufficient statistic	$\bar{X}$-chart
Mean squared error	R-chart
Interval estimation	S-chart
Confidence interval	P-chart

10.1 Introduction

In the previous two chapters, we discussed the basic principles of sampling and sampling distributions – techniques that enable us to make inferences about a population by looking at a subset of that population. In this chapter, we continue our discussion of inferential statistics by examining point estimation, confidence intervals, and statistical quality control. Note that this chapter draws heavily on your understanding of the standard normal distribution discussed in Chap. 7, the fundamental concepts of sampling discussed in Chap. 8, and the t distribution and chi-square distribution discussed in Chap. 9.

We first examine point estimates for population parameters and then discuss desirable attributes of point estimators. Second, basic concepts and the necessity of using interval estimates are discussed in detail. Third, we explain how to compute confidence intervals for population means both when the population variance is known and when it is unknown. Fourth, confidence intervals for the population proportion and the population variance are explored. Finally, we present applications of the use of confidence intervals for quality control. An application of confidence intervals for a cash management model appears in Appendix 1. Appendix 2 shows how MINITAB can be used to generate control charts.

10.2 Point Estimation

As we have said, statistical inference enables us to make judgments about a population on the basis of sample information. The mean, standard deviation, and proportions of a population are called *population parameters*; in other words, they serve to define the population. Estimating a population's parameters is essential to statistical analysis, and sometimes sampling is the best (fastest and most economical) way to approach the study.

10.2.1 Point Estimate, Estimator, and Estimation

A *parameter* is a characteristic of an entire population; a *statistic* is a summary measure that is computed to describe a characteristic for only a sample of the population. An *estimate* is a specific observed value of a statistic. The rule that specifies how a sample statistic can be obtained for estimating the population parameter is called an *estimator*. For example, if a professor wants information on

central tendency in a list of test scores, she can calculate a sample mean. The number for the sample mean is called the estimate, and the sample mean is the estimator for the population mean. The *point estimate* is the single number that is obtained from the estimator.

The symbols we use to represent several important population parameters and their sample counterparts follow.

	Population Parameter	Sample Statistic
Mean	μ	$\bar{X}$
Standard deviation	σ_X	s_X
Variance	σ_X^2	s_X^2
Proportion	p	$\hat{p}$

Example 10.1 Sample Mean and Sample Variance: Point Estimate. Suppose that a professor, whose course has an enrollment of 50 students, wants information on the performance of his class. He takes a sample of 10 scores:

$$95,\ 67,\ 89,\ 70,\ 56,\ 97,\ 68,\ 78,\ 50,\ 79$$

The estimator for the population mean is the sample mean, $\bar{X}$. The estimate for the population mean, on the basis of the 10 sample scores, is $\bar{X} = 74.9$.

The estimator for the population variance is the sample variance. The estimate of the population variance is

$$s_X^2 = \frac{(95^2 + 67^2 + \cdots + 79^2) - 10(74.9)^2}{10 - 1} = 247.65$$

The professor can use $\bar{X} = 74.9$ and $s_X^2 = 247.65$ to do his or her class performance analysis.

The relationship among the point estimate, point estimator, and point estimation can be summarized as follows. A point estimate is a single value that is calculated from only one sample. In Example 10.1, $\bar{X} = 74.9$ is an estimate for population mean μ, and $s_X^2 = 247.65$ is an estimate for population variance σ_X^2. Using the formula for combinations reveals that there are $\binom{50}{10} = 10,272,278,000$ possible sample estimates for Example 10.1.[1] The random variable that is defined by a formula, and from which we obtain all possible estimates, is called the point

[1] From the combination formula discussed in Appendix 1 of Chap. 5, we obtain

$$\binom{50}{10} = \frac{50!}{10!(50-10)!} = \frac{(50)(49)\cdots(41)}{(10)(9)\cdots(1)} = 10,272,278,000$$

Table 10.1 Possible samples and their sample means (sample size = 3)

Possible samples	Elements in sample	Sample mean ($\bar{X}_i$)	Possible samples	Elements in sample	Sample mean ($\bar{X}_i$)
1	1, 2, 3	2	11	2, 3, 4	3
2	1, 2, 4	2.33	12	2, 3, 5	3.33
3	1, 2, 5	2.67	13	2, 3, 6	3.67
4	1, 2, 6	3	14	2, 4, 5	3.67
5	1, 3, 4	2.67	15	2, 4, 6	4
6	1, 3, 5	3	16	2, 5, 6	4.33
7	1, 3, 6	3.33	17	3, 4, 5	4
8	1, 4, 5	3.33	18	3, 4, 6	4.33
9	1, 4, 6	3.67	19	3, 5, 6	4.67
10	1, 5, 6	4	20	4, 5, 6	5
	Sum = 30			Sum = 40	

estimator. A *point estimate* is a single value that is used to estimate a population parameter. A *point estimator* is a sample statistic used to estimate a population parameter. *Point estimation* is a process that generates specific numbers, each of which is a point estimate.

Example 10.2 Population Mean: Point Estimate. We can use a sampling approach to obtain the point estimate of a population mean μ. In Example 8.1, we demonstrated the sampling results of taking samples of 2, 3, or 4 elements out of a uniformly distributed population that represents the numbers of years of working experience of six secretaries (1, 2, 3, 4, 5, and 6) at Francis Engineering Inc. If samples of three elements are randomly taken from this population, then there are 20 possible samples, as listed in Table 10.1.

The population mean and population standard deviation are

$$\mu = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\sigma_X = \left[\frac{(1-3.5)^2 + (2-3.5)^2 + \cdots + (6-3.5)^2}{6}\right]^{1/2} = 1.71$$

All possible sample means listed in Table 10.1 are point estimates of the population mean μ_x. MINITAB output is given in Fig. 10.1, which indicates that both the mean and the median of this set of sample means are equal to 3.5.

10.2.2 Four Important Properties of Estimators

A number of different estimators are possible for the same population parameter, but some estimators are better than others. To understand how, we need to look at four important properties of estimators: unbiasedness, efficiency, consistency, and sufficiency.

```
MTB > SET INTO C1
DATA> 2 2.33 2.67 3 2.67 3 3.33 3.33 3.67
DATA> 4 3 3.33 3.67 3.67 4 4.33 4 4.33 4.67 5
DATA> END
MTB > DESCRIBE C1

Descriptive Statistics

Variable       N     Mean   Median   Tr Mean   StDev   SE Mean
C1            20    3.500    3.500     3.500   0.781     0.715

Variable     Min      Max       Q1       Q3
C2         2.000    5.000    3.000    4.000
```

Fig. 10.1 MINITAB output for Example 10.2

10.2.2.1 Unbiasedness

An estimator exhibits *unbiasedness* when the mean of the sampling estimator $\hat{\theta}$ is equal to the population parameter θ. In other words, the expected value of the estimator is equal to the population parameter: $E(\hat{\theta}) = \theta$. Let's use data given in Table 10.1 as an example:

$$E(\bar{X}_i) = \Sigma P(\bar{X}_i)\bar{X}_i = \frac{\sum_{i=1}^{20} \bar{X}_i}{20} = \frac{2 + 2.33 + \cdots + 4.67 + 5}{20} = \frac{70}{20}$$
$$= 3.5 = \mu$$

Note that $P(\bar{X}_i) = \frac{1}{20}$ because each sample of 3 is equally likely. Figure 10.2 shows the sampling distributions of two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$; $\hat{\theta}_1$ is an unbiased estimator and $\hat{\theta}_2$ a biased estimator. Figure 10.2 indicates that $E(\hat{\theta}_t) = \hat{\theta}$ and $E(\hat{\theta}_2) > \theta$.

In general, unbiasedness is a desirable property for an estimator. The sample mean is an unbiased estimator of the population mean because the mean of the sampling distribution of $\bar{X}$, $E(\bar{X})$, is equal to the population mean μ. Similarly, the sample variance is an unbiased estimator of the population variance because the mean of the sample distribution of s_X^2, $E(s_X^2)$, is equal to population variance σ_X^2.[2] And the sample proportion is an unbiased estimator of the population proportion; $E(\hat{p}) = p$. However, because standard deviation is a *nonlinear* function of variance, the sample standard deviation is not an unbiased estimator of population standard deviation.

The *bias* of a point estimator is defined in Eq. 10.1:

$$\text{Bias} = \text{E}(\hat{\theta}) - \theta \tag{10.1}$$

[2] If we divide the sum of squared discrepancies from $\bar{X}$ by $(n-1)$ rather than n, Eq. 9.11 in Chap. 9 can be used to demonstrate this point.

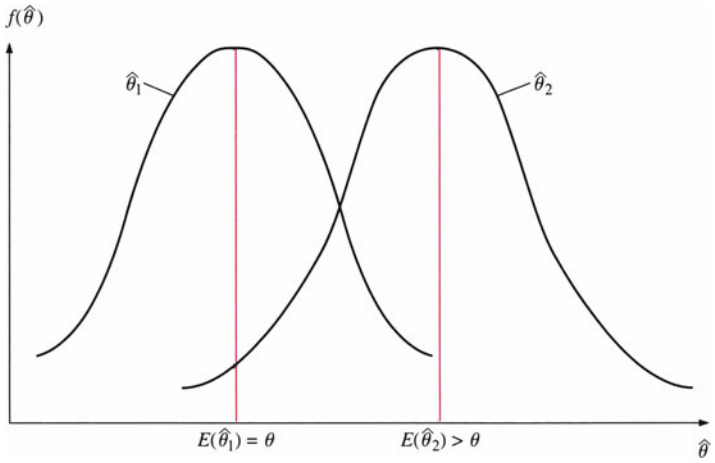


Fig. 10.2 Probability density functions for θ_1 and θ_2

For example, in Fig. 10.2 the bias of using $\hat{\theta}_2$ as an estimator of θ is equal to $E\ (\hat{\theta}_2) - \theta$.

Unbiasedness, then, is an important attribute of estimators. But suppose we have a number of unbiased estimators to choose from. Here are three other criteria that could be used to select an estimator.

10.2.2.2 Efficiency

Efficiency is another standard that can be used to evaluate estimators. *Efficiency* refers to the size of the standard error of the statistics. The most efficient estimator is the one with the smallest variance. Thus, if there are two estimators for θ with variances $\text{Var}(\hat{\theta}_1)$ and $\text{Var}(\hat{\theta}_2)$, then the first estimator $\hat{\theta}$ is said to be more efficient than the second estimator $\hat{\theta}_2$ if $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$ although $E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$. Figure 10.3 shows the distributions of the two density functions.

The relative efficiency of one estimator compared with another is simply the ratio of their variances. Given two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ with variances $\text{Var}(\hat{\theta}_1)$ and $\text{Var}(\hat{\theta}_2)$, the relative efficiency of $\hat{\theta}_2$ with respect to $\hat{\theta}$ is

$$\text{Relative efficiency} = \frac{\text{Var}(\hat{\theta}_1)}{\text{Var}(\hat{\theta}_2)} \tag{10.2}$$

Why is the variance of the benchmark estimator $(\hat{\theta}_1)$ placed in the numerator? Well, suppose two estimators are calculated for the population mean. The first is the sample mean $\hat{\theta}_1$ and the second is the sample median $\hat{\theta}_2$. It can be shown that the

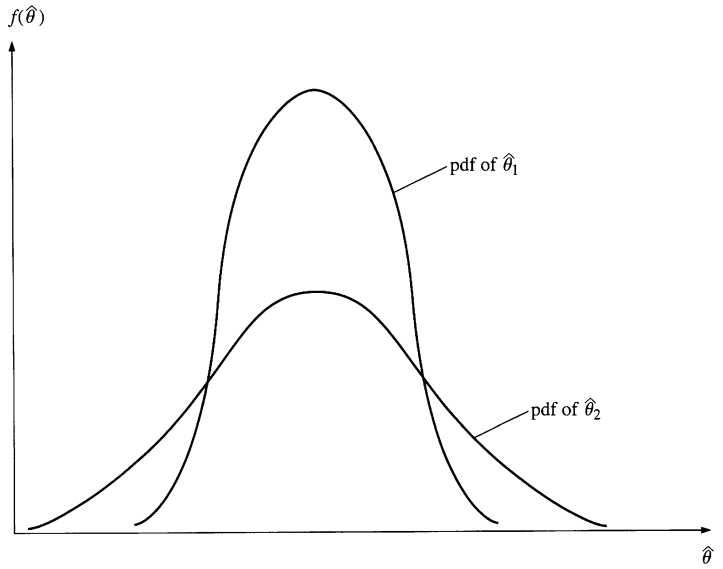


Fig. 10.3 Probability density functions of two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$; $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$

variance of the sample median of a normal distribution is $\text{Var}(\hat{\theta}_2) = \pi(\sigma_X^2/2n)$. The variance for the sample mean is σ_X^2/n. The relative efficiency of the sample median with respect to the sample mean is

$$\text{Efficiency} = \frac{\text{Var}(\hat{\theta}_1)}{\text{Var}(\hat{\theta}_2)} = \frac{\sigma_X^2/n}{\pi\,\sigma_X^2/2n} = \frac{2}{\pi} = 63.66\%$$

The sample mean, rather than the sample median, is the preferred estimator of the population mean because the amount of variability associated with the sample mean is about 64 % of that associated with the sample median. Note that the sample mean is the best estimate of central tendency for symmetric distributions and that the sample median is generally used for skewed distributions.

10.2.2.3 Consistency

A third property of estimators, *consistency*, is related to their behavior as the sample size gets large. A statistic is a consistent estimator of a population parameter if, as the sample size increases, it becomes almost certain that the value of the statistic comes very close to the value of the population parameter. For example, suppose we are tossing a coin and are interested in rolling a head. The sample proportion X/n is an estimator for the population proportion, where X is the number of heads tossed and n is the number of trials. We know that the population proportion of heads tossed is equal to 1/2, so we would expect the sample proportion to get closer to 1/2 as the

number of trials n increases. (This result was demonstrated by a computer simulation in Chap. 5.) We need information on the probability that the absolute difference between the estimator and the parameter will be less than some positive number ϵ.

In other words, we need $P(|X/n - p| \leq \epsilon)$, and this probability should be close to 1 as n gets large. If it is, then X/n is said to be a consistent estimator of p.

It can be shown that an unbiased estimator $\hat{\theta}_n$ for θ is a consistent estimator if the variance approaches 0 as n increases. For example, we can show that the sample mean is a consistent estimator of the population. The sample mean is unbiased because $E(\bar{X}) = \mu$. The variance of X is σ_X^2/n. As n becomes large, the variance gets closer to 0; this estimator is consistent. Finally, it should be noted that the sample standard deviation is a consistent estimator of population standard deviation, although it is not an unbiased estimator of population standard deviation.

Following this approach, we can see that the sample proportion $X/n = \hat{p}$ is also consistent. From the last chapter, we know that $E(\hat{p}) = p$, which establishes unbiasedness. Because the variance is equal to $\hat{p}(1 - \hat{p})/n$, the variance approaches 0 as n gets large. Thus, X/n is a consistent estimator of p. $\hat{\theta}_n$ is a consistent estimator of θ if, for any positive number ϵ

$$\lim_{n \to \infty} P\left(\left|\hat{\theta}_n - \theta\right| \leq \epsilon\right) = 1 \tag{10.3}$$

where $n \to \infty$ means that sample size approaches infinity.

10.2.2.4 Sufficiency

The last property of a good estimator that we will consider is sufficiency, which was developed by Sir R. A. Fisher, a famous statistician, in 1922.[3] A *sufficient statistic* (such as $\bar{X}$) is an estimator that utilizes all the information a sample contains about the parameter to be estimated. For example, $\bar{X}$ is a sufficient estimator of the population mean μ. This means that no other estimator of μ from the same sample data, such as the sample median, can add any further information about the parameter μ that is being estimated.

It can be shown that the sample mean $\bar{X}$ and the sample proportion $\hat{p}$ are sufficient statistics (estimators) for μ and p.

10.2.3 Mean Squared Error for Choosing Point Estimator

Frequently, a trade-off must be made between bias and efficiency for a point estimator. Sometimes there is much to be gained by accepting some biases for the sake of increasing the efficiency of an estimator. A statistic called the *mean squared*

[3] R. A. Fisher (1922), "On the Mathematical Foundations of Theoretical Statistics," *Phil. Trans. Roy. Soc. London*. Series A, Vol. 222.

error (MSE), the expectation of the squared difference between the estimators and parameters as indicated in Eq. 10.4, can be used to measure the trade-off between bias and efficiency for an estimator:

$$\mathrm{MSE}\left(\hat{\theta}\right) = E\left(\hat{\theta} - \theta\right)^2 \tag{10.4}$$

It can be shown[4] that

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + [\mathrm{Bias}(\hat{\theta})]^2 \tag{10.5}$$

where $\mathrm{Var}(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$ and $\mathrm{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Equations 10.4 and 10.5 imply that an estimator's variance is a measure of the dispersion of the sampling distribution around the estimator's expected value, $E(\hat{\theta})$, whereas the MSE is a measure of dispersion around the true population parameters, θ. If the estimator is unbiased, then $E(\hat{\theta}) = \theta$ and $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\theta)$. For example, in Fig. 10.3 the expected variability of $\hat{\theta}_2$ around the true parameter, θ, is greater than that around $E(\hat{\theta}_2)$, which is the center of the sample distribution. Both nonsampling error and systematic sampling error bias an estimator.

10.3 Interval Estimation

In the last section, we discussed point estimation of a population parameter. We investigated methods for estimating population mean, variance, standard deviation, and proportion and methods for evaluating desirable features of estimators. Although these estimators give us much information about a population parameter, more information is usually desired. Many times, an interval estimate is needed. For example, a manager may want to know how likely it is that the mean number of defects is between 1 % and 3 %, or a professor may want to know how likely it is that between 10 % and 20 % of her class will get an A on the final exam. Sample statistics such as the mean and variance do not provide any information on the range of values the population parameters are likely to fall in.

We now wish to estimate a parameter μ by the interval

$$a < \mu < b$$

4

$$\begin{aligned} E(\hat{\theta} - \theta)^2 &= E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 + 2E[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta] \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 \quad \text{because } E[\hat{\theta} - E(\hat{\theta})] = 0 \end{aligned}$$

where a and b are obtained from sample observation. The estimation of a and b, values between which the parameter of interest will lie with a certain probability, is called *interval estimation*.

Suppose θ is a parameter to be estimated. A random sample is taken, and two random variables a and b are computed. The interval from a to b is called a *confidence interval*; its probability is $(1-\alpha)$. In other words, if all of the population is repeatedly sampled and the intervals are calculated in the same fashion, then the probability is $(1-\alpha)$ that the confidence interval will contain the population parameter:

$$P(a < \theta < b) = 1 - \alpha \tag{10.6}$$

For example, if $1-\alpha = .95$, then the probability that a is less than θ and b is greater than θ is .05. Because $1-\alpha = .95$, $\alpha = .05$. The term $(1-\alpha)$ is called the *confidence level* (*probability content*). The quantity α is often termed the *risk probability* or *significant level*. In the next four sections, we will use Eq. 10.6 to estimate the confidence interval for population mean, population proportion, and population variance.

10.4 Interval Estimates for μ When σ_X^2 Is Known

In this section, we construct confidence intervals for the population mean. We assume that the random sample is taken from a normal distribution and that the population variance is known. The latter assumption is somewhat unrealistic because the population variance is rarely known. However, these assumptions enable us to illustrate concepts that we will need later.

Suppose a random sample is taken with an unknown mean and known variance. The confidence interval uses the fact that the random variable Z where

$$Z = \frac{\bar{X} - \mu}{\sigma_X / \sqrt{n}}$$

has a standard normal distribution. Suppose a $100(1-\alpha)$ percent confidence interval is set up, so that $\alpha/2$ is the area of the right tail of the normal distribution, $\alpha/2$ is the area of the left tail, and $(1-\alpha)$ is the area in the center, as shown in Fig. 10.4. The cutoff points on the normal distribution are $z_{\alpha/2}$ and $-z_{\alpha/2}$. The confidence interval is derived as follows:

$$\begin{aligned} 1 - \alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &= P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma_X / \sqrt{n}} < z_{\alpha/2}\right) \end{aligned}$$

$$1 - \alpha = P\left\{\bar{X} - z_{\alpha/2}\left[\frac{\sigma_X}{\sqrt{n}}\right] < \mu < \bar{X} + z_{\alpha/2}\left[\frac{\sigma_X}{\sqrt{n}}\right]\right\} \tag{10.7}$$

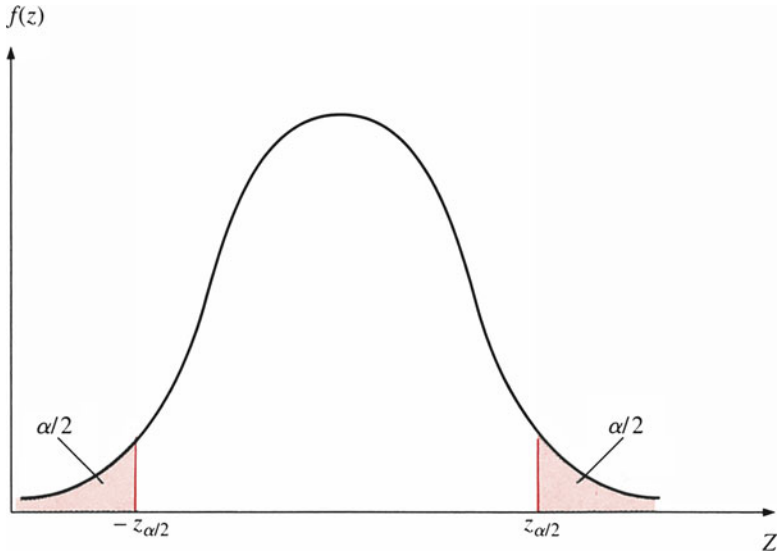


Fig. 10.4 Risk probability for sample mean estimate

Equation 10.7 implies that confidence intervals have the following characteristics:

1. As the standard deviation increases, the length of the confidence interval increases. This result is understandable: the wider the deviation, the more uncertain the estimate of the mean.
2. The bigger the sample size, the smaller the confidence interval for a given variance. This is because more information decreases the interval, making a better interval possible.
3. The confidence interval is larger for smaller confidence levels (α). A 99 % confidence interval has a smaller α than a 95 % interval because a 99 % interval has more certainty.

Example 10.3 Confidence Intervals in Terms of 20 Samples. Let us now refer to the 20 samples in Table 10.1 and use the information $\sigma_x = 1.71$ and the 20 random samples given there. We calculate 20 different 95 % confidence intervals in terms of Eq. 10.7 by using MINITAB as presented in Fig. 10.5. The 95 % confidence interval (CI) results listed in Fig. 10.5 reveal that all 20 samples resulted in a confidence interval containing $\mu = 3.5$.

```
MTB >  READ   C1-C20
DATA>  1  1  1  1   1  1   1  1   1   1  2  2  2  2  2  3  3  3  4
DATA>  2  2  2  2   3  3   3  4   4   5  3  3  3  4  4  5  4  4  5  5
DATA>  3  4  5  6   4  5   6  5   6   6  4  5  6  5  6  6  5  6  6  6
DATA>  END
       3 rows read.
MTB >  PRINT C1-C20.
```

Data Display

Row	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
2	2	2	2	2	3	3	3	4	4	5	3	3	3	4
3	3	4	5	6	4	5	6	5	6	6	4	5	6	5

Row	C15	C16	C17	C18	C19	C20
1	2	2	3	3	3	4
2	4	5	4	4	5	5
3	6	6	5	6	6	6

```
MTB >  STORE
STOR>  ZINTERVAL USING 95%, SIGMA=1.71, DATA IN CK1
STOR>  LET K1=K1+1
STOR>  END
MTB >  LET K1=1
MTB >  EXECUTE 'MINITAB' 20
Executing from file: MINITAB.MTB
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE Mean          95.0  %  CI
C1             3     2.000      1.000      0.987       (0.065,    3.935)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE Mean          95.0  %  CI
C2             3     2.333      1.528      0.987       (0.398,    4.269)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE Mean          95.0  %  CI
C3             3     2.667      2.082      0.987       (0.731,    4.602)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE Mean          95.0  %  CI
C4             3     3.000      2.646      0.987       (1.065,    4.935)
```

Fig. 10.5 (continued)

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE Mean          95.0  %  CI
C5             3     2.667      1.528      0.987       (0.731,    4.602)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE Mean          95.0  %  CI
C6             3     3.000      2.000      0.987       (1.065,    4.935)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev   SE  Mean          95.0  %  CI
C7             3     3.333      2.517      0.987       (1.398,    5.269)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE Mean          95.0  %  CI
C8             3     3.333      2.082      0.987       (1.398,    5.269)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev   SE  Mean          95.0  %  CI
C9             3     3.667      2.517      0.987       (1.731,    5.602)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev   SE  Mean          95.0  %  CI
C10            3     4.000      2.646      0.987       (2.065,    5.935)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev   SE  Mean          95.0  %  CI
C11            3     3.000      1.000      0.987       (1.065,    4.935)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev   SE  Mean          95.0  %  CI
C12            3     3.333      1.528      0.987       (1.398,    5.269)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev   SE  Mean          95.0  %  CI
C13            3     3.667      2.082      0.987       (1.731,    5.602)
```

Fig. 10.5 (continued)

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE  Mean        95.0  %  CI
C14            3      3.667     1.528      0.987   (1.731,          5.602)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE  Mean        95.0  %  CI
C15            3      4.000     2.000      0.987   (2.065,        5.935)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE  Mean        95.0  %  CI
C16            3      4.333     2.082      0.987   (2.398,        6.269)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE  Mean        95.0  %  CI
C17            3      4.000     1.000      0.987   (2.065,        5.935)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE  Mean        95.0  %  CI
C18            3      4.333     1.528      0.987   (2.398,        6.269)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE  Mean        95.0  %  CI
C19            3      4.667     1.528      0.987   (2.731,        6.602)
```

Confidence Intervals

```
The assumed sigma = 1.71

Variable       N      Mean      StDev    SE  Mean        95.0  %  CI
C20            3      5.000     1.000      0.987   (3.065,        6.935)
```

Fig. 10.5 MINITAB output for Example 10.3

Example 10.4 Sandbags We Can Have Real Confidence In: 95 % and 99 % Confidence Intervals. Suppose a machine dispenses sand into bags. The population standard deviation is 9.0 lb, and the weights are normally distributed. A random sample of 100 bags is taken, and the sample mean is 105 lb. Let's calculate a 95 % confidence interval by using Eq. 10.7.

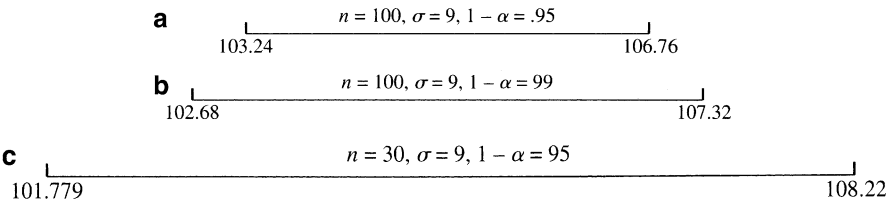


Fig. 10.6 (**a**) and (**b**) The effects of sample size and probability content on confidence intervals in cases of sample mean of 105 (**c**) The effects of sample size and probability content on confidence interval in the case of sample mean of 105

The estimate of sample standard deviation is $9/\sqrt{100}$. The confidence level is $(1-\alpha) = .95$. Thus, $\alpha = .05$ and $\alpha/2 = .025$. The Z-value that corresponds to this area is 1.96. This is due to the fact that the area of the right tail is .025. The area of the left tail is .025. The 95 % confidence interval is

$$105 - (1.96)\frac{(9)}{\sqrt{100}} < \mu < 105 + (1.96)\frac{(9)}{\sqrt{100}}$$

$$\text{or} \qquad 103.24 < \mu < 106.76$$

The 95 % confidence interval for the mean weight of the bags ranges from 103.24 to 106.76, as presented in Fig. 10.6a. In other words, we are certain that 95 % of such intervals contain the population mean.

Suppose a 99 % confidence interval is needed for the case discussed here. The confidence level is $1-\alpha = .99$. Thus, $\alpha = .01$ and $\alpha/2-.005$. The Z-value is 2.575. The confidence interval is

$$105 - (2.575)\frac{(9)}{\sqrt{100}} < \mu < 105 + (2.575)\frac{(9)}{\sqrt{100}}$$

or

$$102.68 < \mu < 107.32$$

This 99 % confidence interval is wider than the 95 % confidence interval, as indicated in Fig. 10.6b. Now we are certain that the interval generated by our statistics will include the true population mean 99 % of the time.

Example 10.5 95 % Confidence Interval for the Sandbag Sample with a Smaller Sample Size. Assume that rather than taking a sample of 100, we take a sample of 30. The 95 % confidence interval becomes

$$105 - (1.96)\frac{(9)}{\sqrt{30}} < \mu < 105 + (1.96)\frac{(9)}{\sqrt{30}}$$

or

$$101.779 < \mu < 108.22$$

This interval is wider than the interval with a sample size of 100, as indicated in Fig. 10.6c.

Example 10.6 95 % Confidence Interval for the Mean External Audit Fees for 32 Diverse Companies. To study the effect of internal audit departments on external audit fees, W. A. Wallace recently conducted a survey of the audit departments of 32 diverse companies (*Harvard Business Review*, March–April 1984). She found that the mean annual external audit paid by the 32 companies was $779,030 and the standard deviation was $1,083,162.

Because this is a large-sample case, we can replace the sample standard deviation for the population standard deviation. Substituting both the sample mean and the sample standard deviation and other information into Eq. 10.7, we obtain the 95 % confidence interval as

$$779,030 - (1.96)\frac{(1,083,162)}{\sqrt{32}} < \mu < 779,030 + (1.96)\frac{(1,083,162)}{\sqrt{32}}$$

or

$$403,733.23 < \mu < 1,154,326.77$$

A 95 % confidence interval for the mean external audit fees paid by all companies during the year ranges from $403,733.23 to $ 1,154,326.77.

10.5 Confidence Intervals for μ When σ_X^2 Is Unknown

In the previous section, we constructed confidence intervals for known population variances. For a large sample size, the assumption of known population variance can be relaxed. In this section, we construct confidence intervals for small sample sizes ($n < 30$) and unknown population variance.

In some cases, it is not possible to obtain a large sample size. For example, we might be interested in constructing a confidence interval for sales in a particular industry that contains only 10 firms. Because of the size of the sample, the normal distribution cannot be used; the central limit theorem applies only to large sample sizes ($n \geq 30$).

Remember that the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma_X / \sqrt{n}}$$

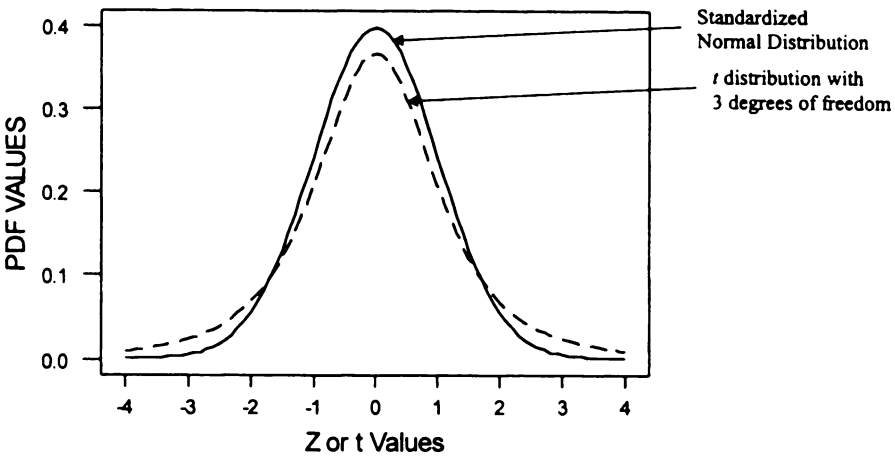


Fig. 10.7 Standardized normal distributions versus t distribution

has a standard normal distribution. However, in our example the sample size is small and the population standard deviation is unknown, so the t statistic discussed in Chap. 9 must be used. Recall that the t statistic has a shape that is very similar to the normal distribution, which is bell-shaped and symmetrically distributed.

However, the t distribution has fatter tails than the Z (the standardized normal) distribution. This is because using the sample standard deviation rather than the population standard deviation introduces uncertainty. The similarities and differences between the two distributions are shown in Fig. 10.7.

The probability in the tail and the shape of the distribution depend on the number of degrees of freedom, $n-1$, where n is the number of observations in the sample. Table A4 in Appendix A shows the relationship between the number of degrees of freedom and the t value of the distribution. For example, if the number of degrees of freedom equals 6 and the area in both tails combined is 0.100, then the t value is 1.943. If the degrees of freedom equal 12, then the t value is 1.782. In addition, the t value will be greater than the Z-value for the same area under the curve. As the number of degrees of freedom approaches infinity, the t distribution approaches the Z distribution. This is due to the fact that as n becomes larger, the sample standard deviation s approaches the population standard deviation σ. We use a cutoff $n = 30$ to distinguish between large and small samples because there is little difference between the two distributions at that sample size.

The random variable t with $v = (n-1)$ degrees of freedom can be defined as

$$t_v = \frac{\bar{X} - \mu}{s_X/\sqrt{n}} \tag{10.8}$$

which follows the t distribution with $(n-1)$ degrees of freedom. Note that this is similar to the Z statistic, but the sample standard deviation is used instead of the population standard deviation.

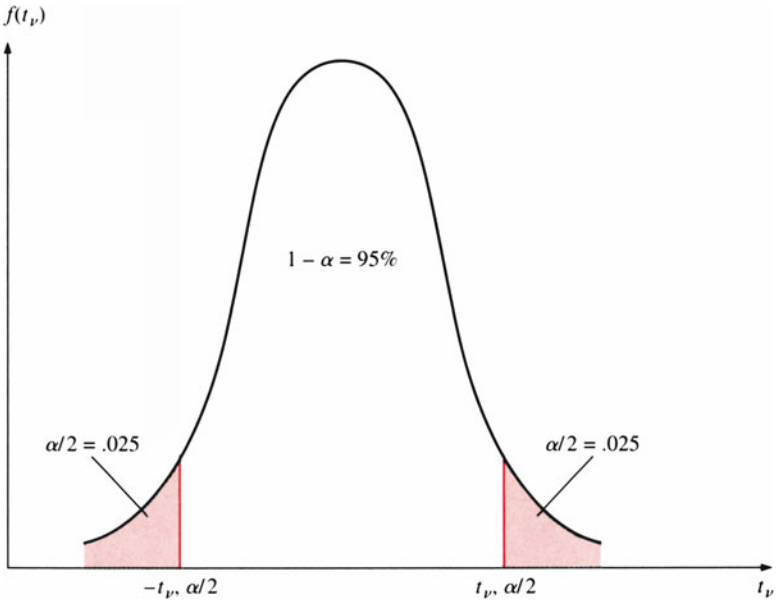


Fig. 10.8 A t distribution with $n = 10$ and $1-\alpha = 95$ %

To use the *t* distribution for confidence intervals, we need to take areas on both sides of the distribution. Thus, if we want a 95 % confidence interval with a sample size of 10, we divide 5 % by 2 (0.05/2 = 0.025) and look up 0.025 with (10–1) degrees of freedom in the *t* tables to arrive at a *t* value. The positive value will correspond to the right-side tail and the negative value to the left-side tail. This is shown in Fig. 10.8.

Using the *t* distribution, we find that the confidence interval is

$$1-\alpha = P\left[-t_{n-1,\alpha/2} < \frac{\bar{X}-\mu}{s_X/\sqrt{n}} < t_{n-1,\alpha/2}\right] = P\left[\bar{X} - t_{n-1,\alpha/2}\frac{s_X}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,\alpha/2}\frac{s_X}{\sqrt{n}}\right] \quad (10.9)$$

Example 10.7 95 % Confidence Interval for the Average Weight of Football Players. A random sample yields the following weights of eight football players, in pounds.

250 210 185 242 190 200 220 205

The sample mean is $\bar{X} = 212.75$, and the sample standard deviation is $s_X = 23.34$.

Suppose we want a 95 % confidence interval. The number of degrees of freedom is 8–1 = 7, and the corresponding *a* is 0.05. The *t* value that is desired is $t_{n-1,a/2} = t_{7,0.05/2} = 2.365$.

```
MTB >    SET INTO C1
DATA>    250  210  185  242  190  200  220  205
DATA>    END
MTB >    TINTERVAL WITH  95%  CONFIDENCE  USING  C1

Confidence Intervals

Variable    N    Mean    StDev  SE  Mean      95.0  %  CI
C1          8    212.75  27.34   8.25    (193.22 ,   232.28)
```

Fig. 10.9 MINITAB solution to Example 10.7

Substituting all this information into Eq. 10.7, we get

$$212.75 - (2.365)\frac{(23.34)}{\sqrt{8}} < \mu < 212.75 + (2.365)\frac{(23.34)}{\sqrt{8}}$$

or

$$193.23 < \mu < 232.27$$

Thus, we can say with 95 % certainty that the true mean weight of the football players is between 193.2 and 232.3 lb. The MINITAB solution for this example is shown in Fig. 10.9.

Example 10.8 90 % Confidence Interval for the Average Weight of Football Players. Suppose a 90 % confidence interval is constructed for the information given in Example 10.7. The t value is $t_{7,\,0.10/2}$. The 90 % confidence interval is

$$212.75 - (1.895)\frac{(23.34)}{\sqrt{8}} < \mu < 212.27 + (1.895)\frac{(23.34)}{\sqrt{8}}$$

or

$$197.11 < \mu < 228.39$$

Here, because we have chosen a 90 % confidence interval, the confidence interval has gotten narrower.

Example 10.9 Estimate for Waiting Time at a Bank. As part of an effort to improve customer service, a bank pledges not to keep customers waiting in line an unreasonable time. To determine the time interval of waiting in line, the bank collects the following data for nine customers.

Customer	Waiting time (min)
A	4
B	3
C	6
D	2
E	7
F	1
G	3
H	4
I	2

The mean is $\bar{X} = 3.56$, and the standard deviation is $s_x = 1.94$. The t value is $t_{8,\ 0.05/2} = 2.306$. The bank constructs a 95 % confidence interval for the mean waiting time per customer. It is

$$3.56 - (2.306)\frac{(1.94)}{\sqrt{9}} < \mu < 3.56 + (2.306)\frac{(1.94)}{\sqrt{9}}$$

or

$$2.069 < \mu < 5.051$$

The bank concludes that the true mean number of minutes a customer must wait is between 2.069 and 5.051 min with 95 % probability.

Example 10.10 95 % Confidence Interval for the True Mean Incremental Profit of "Successful" Trade Promotion. Each year, thousands of manufacturers' sales promotions are conducted by North American packaged goods companies. A sample of Canadian packaged goods companies provided information on examples of past sales promotion, including trade promotion. By interviewing the company managers, K. G. Hardy (*Journal of Marketing*, July 1986, Vol. 50, No. 7) identified 21 "successful" sample trade promotions with the mean incremental profit $53,000 and the standard deviation $95,000.

If the population from which the sample is selected has an approximate normal distribution, then the 95 % confidence interval for the true mean incremental profit of "successful" trade promotion can be calculated in terms of Eq. 10.9 as

$$53,000 - (2.086)\frac{(95,000)}{\sqrt{21}} < \mu < 53,000 + (2.086)\frac{(95,000)}{\sqrt{21}}$$

or

$$9,755.99 < \mu < 96,244.01$$

Hardy concluded that the true mean incremental profit of "successful" trade promotions is between $9,755.99 and $96,244.01 with 95 % probability.

10.6 Confidence Intervals for the Population Proportion

Suppose a quality control expert needs to determine the proportion of defective parts for a company – that is, the proportion of a particular item that is returned by the company's customers. Or suppose a political analyst would like to report the proportion of voters who support a particular candidate for a US Senate race. In this section we will derive confidence intervals for population proportions. The concepts are similar to those used in the section on large-sample mean confidence intervals because the standard normal distribution is used in both.

In Chap. 8 we found that for large sample sizes, the random variable

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \tag{10.10a}$$

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \tag{10.10b}$$

has a normal distribution, where $\hat{p}$ and p are the sample proportion and the population proportion, respectively. Equation 10.10a is defined in terms of population standard deviation, and Eq. 10.10b is defined in terms of sample standard deviation. We will use Eq. 10.10b to develop a confidence interval for the population proportion:

$$\begin{aligned} 1 - \alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &= P\left[-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} < z_{\alpha/2}\right] \end{aligned}$$

where $z_{\alpha/2}$ is the number such that $P(Z > z_{\alpha/2}) = \alpha/2$. We now move all the terms except the population proportion to the right and left sides of P, which gives a $(1-\alpha)$ confidence interval.

$$1 - \alpha = P\left\{\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right\} \tag{10.11}$$

In order for us to use the Z statistic, the sample size must be large. For most purposes, a sample size that is greater than 30 will do. By looking at the confidence interval, we can see that the larger the α value, the smaller the Zs (and the confidence interval) will be. In addition, as the sample size increases, the confidence interval gets narrower.

Example 10.11 95 % Confidence Interval for Voting Proportion. Suppose that a random sample of 100 voters is taken, and 55 % of the sample supports the

incumbent candidate. Construct a 95 % confidence interval for this proportion. The sample size is large, so we can use Eq. 10.11 to obtain the interval.

The Z-value for a 95 % confidence interval is $z_{0.05/2} = 1.96$.

$$.55 - 1.96\sqrt{\frac{.55(1-.55)}{100}} < p < .55 + 1.96\sqrt{\frac{.55(1-.55)}{100}}$$

or

$$.452 < p < .648$$

The 95 % confidence interval for the true proportion of voters supporting the incumbent goes from 45.2 % to 64.8 %.

Now suppose we want a 90 % confidence interval.

$$.55 - 1.645\sqrt{\frac{.55(1-.55)}{100}} < p < .55 + 1.645\sqrt{\frac{.55(1-.55)}{100}}$$

or

$$.468 < p < .632$$

As we have come to expect, the 95 % confidence interval is wider than the 90 % confidence interval.

Example 10.12 95 % Confidence Interval for Commodity Preference Proportion. A marketing firm discovers that 65 % of the 30 customers who participated in a blind taste test prefer brand A than brand B. The firm develops a 95 % confidence interval in terms of Eq. 10.10b for the number of people who prefer brand A.

$$.65 - 1.96\sqrt{\frac{.65(1-.65)}{30}} < p < .65 + 1.96\sqrt{\frac{.65(1-.65)}{30}}$$

or

$$.479 < p < .821$$

The firm can be 95 % certain that the true proportion of those who prefer brand A lies between 47.9 % and 82.1 %. If the sample size were increased, the confidence interval would be narrower.

Example 10.13 95 % Confidence Interval for the Proportion of Working Adults Who Use Computer Equipment. A recent study (*Journal of Advertising Research*, April/May 1984) to find the proportion of working adults using computer equipment (personal computers, microcomputers, computer terminals, or word processors) on

the job employed the random sample approach to survey 616 working adults. The survey revealed that 184 of the adults now regularly use computer equipment on the job.

A 95 % confidence interval for working adults' computer usage can be calculated in accordance with Eq. 10.11. In this case, $n = 616$; $p = 184/616 = .299$; $z_{.05/2} = 1.96$.

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(.299)(.701)}{616}} = .018$$

Substituting all of this information into Eq. 10.11, we obtain

$$.299 - (1.96)(.018) < p < .299 + (1.96)(.018)$$

or

$$.264 < p < .334$$

In other words, the 95 % confidence interval for the true proportion of all working adults who regularly use computer equipment on the job is between 0.264 and 0.334.

10.7 Confidence Intervals for the Variance

Despite an increased awareness of the importance of quality and despite the subsequent introduction of robots and other precision tools into factories, some variance is inevitable in any manufacturing process. Manufacturers need to know whether the variance falls within an acceptable range. To determine this, they construct confidence intervals. We have seen that confidence intervals can be constructed by using the normal distribution (large-sample population means) and the *t* distribution (small-sample population means). For variance we must use the chi-square distribution because, as we noted in Chap. 9, the variance is χ^2 distributed.

Figure 10.10 shows the chi-square distribution and its confidence interval. The area of the middle part is $(1-\alpha)$, the area of the right tail is $\alpha/2$, and the area of the left tail is $\alpha/2$. The number corresponding to the right tail is $\chi^2_{\nu,\,\alpha/2}$ and the number for the left tail is $\chi^2_{\nu,1-\alpha/2}$, where ν is the degrees of freedom (the number in the sample less one, $n-1$). For example, if a 90 % confidence interval is desired with a sample size of 20, then the critical values are $\chi^2_{19,.05} = 30.1435$ and $\chi^2_{19,.95} = 10.1170$.

As discussed in Chap. 9, the random variable

$$\chi^2_\nu = \frac{(n-1)s^2_X}{\sigma^2_X}$$

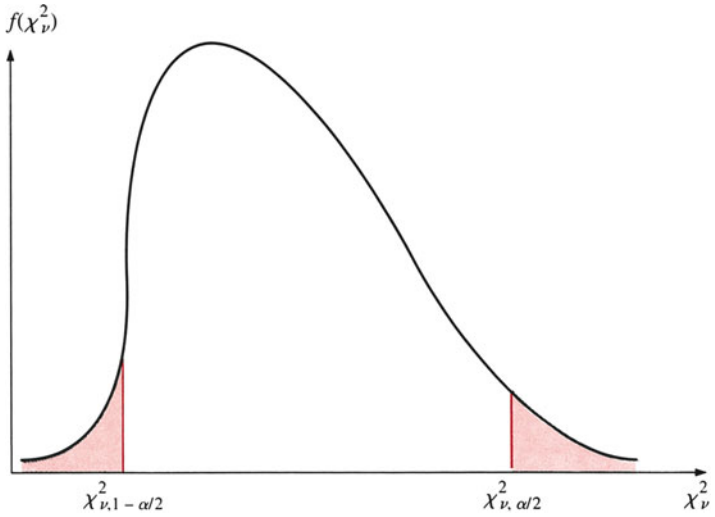


Fig. 10.10 χ^2 distribution

is a chi-square random variable with $v = (n-1)$ degrees of freedom. The confidence interval for the population variance is derived as follows:

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{v,1-\alpha/2} < \chi^2_v < \chi^2_{v,\alpha/2}\right) \\ &= P\left(\chi^2_{v,1-\alpha/2} < \frac{(n-1)s_X^2}{\sigma_X^2} < \chi^2_{v,\alpha/2}\right) \\ &= P\left(\frac{(n-1)s_X^2}{\chi^2_{v,\alpha/2}} < \sigma_X^2 < \frac{(n-1)s_X^2}{\chi^2_{v,1-\alpha/2}}\right) = 1-\alpha \end{aligned} \tag{10.12}$$

Hence, if s_X^2 is the sample variance estimate, it follows that a $100(1-\alpha)$ percent confidence interval for a population variance is given by

$$\frac{(n-1)s_X^2}{\chi^2_{v,\alpha/2}} < \sigma_X^2 < \frac{(n-1)s_X^2}{\chi^2_{v,1-\alpha/2}} \tag{10.13}$$

This formula gives us a $(1-\alpha)$ confidence interval for the population variance. We find a confidence interval for the standard deviation simply by taking the square root of the upper and lower limits.

Example 10.14 Confidence Intervals for σ_X^2. Suppose a random sample of 30 bags of sand is taken and the sample variance of weight is 5.5. Find a 95 % confidence interval for the population variance of the bags. In this example, $\alpha = 0.05$ and

$$\chi^2_{v,\alpha/2} = \chi^2_{29,.025} = 45.7222 \quad \text{and} \quad \chi^2_{v,1-\alpha/2} = \chi^2_{29,.975} = 16.0471$$

Substituting all related information into Eq. 10.13, we obtain

$$\frac{(29)(5.5)}{45.7222} < \sigma_X^2 < \frac{(29)(5.5)}{16.0471}$$

or

$$3.488 < \sigma_X^2 < 9.94$$

This implies that the 95 % confidence interval for the sample variance of sandbag weight is between 3.488 and 9.94.

10.8 An Overview of Statistical Quality Control[5]

Consumers are generally looking for a product that offers reasonable quality at a reasonable price. The quality of a good or service is often perceived by the consumer in terms of appearance, operation, and reliability. Examples of these three dimensions are listed in Table 10.2. Therefore, product or service quality should generally be managed and controlled in accordance with these criteria.

Stevenson, Grant and Leavenworth, Griffith, Evans and Lindsay, and others have shown that statistical methods are key ingredients for the management and control of product or service quality.[6] Three basic statistical quality control issues are:

How much to inspect and how often Acceptance sampling Process control

The first two issues involve determination of the sample size and the sampling methods used for statistical quality control, which will be discussed in this section. Process control consists of (1) the construction and application of control charts in doing quality control and (2) related statistical analysis and testing of control charts. The construction and application of control charts will be discussed in the next section. Further statistical analysis and testing of control charts will be discussed in Chap. 11.

[5] This and the next section are essentially drawn from J. R. Evans and W. M. Lindsay (1989), *The Management and Control of Quality* (St. Paul, MN: West), Chaps. 12, 13, and 15. Reprinted by permission by West Publishing Company. All rights reserved. The main reason for including quality control in this chapter is that the construction and use of control charts in process control are similar to the construction and use of interval estimates discussed in the last five sections. Note, however, that the interval estimate focused on the static estimate of confidence intervals based on fixed populations, whereas quality control charts involve the dynamic estimate of confidence intervals to detect potential changes in populations.

[6] W. J. Stevenson (1990), *Production/Operations Management*, 3rd ed. (Homewood, IL: Irwin); E. L. Grant and R. S. Leavenworth (1988), *Statistical Quality Control*, 6th ed. (New York: McGraw-Hill); G. K. Griffith (1989), *Statistical Process Control Methods for Long and Short Runs* (*Milwaukee*, WI: ASQC Quality Press); and J. R. Evans and W. M. Lindsay (1989), *The Management and Control of Quality*, (St. Paul, MN: West).

Table 10.2 Examples of dimensions of quality

Product/ service	Appearance	Operation	Reliability
Color TV	Cabinetry, position of controls, exterior workmanship	Clarity, sound, ease of adjustment, reception, realistic colors	Frequency of repair
Clothing	Seams matched, no loose threads or missing buttons, pattern matched, fit, style	Warm/cool, resistance to wrinkles, colorfastness	Durability
Restaurant meal	Color, arrangement, atmosphere, cleanliness, friendliness of servers	Taste and consistency of food	Indigestion?

Source: Stevenson (1990), Table. 16.1, p. 808

10.8.1 The Sample Size of an Inspection

The amount of inspection can range from conducting no inspection at all to scrutinizing each item many times. Low-cost, high-volume items such as paper clips, paper cups, and wooden rulers often require little inspection because the cost associated with defectives is low and the processes of production are usually very reliable. On the other hand, high-cost, low-volume items such as critical components of an occupied space vehicle are closely scrutinized because of the risk to human safety and high cost of mission failure. The majority of quality control applications lie somewhere between these two extremes, and here sampling comes into play.

The sample size of sampling surveys is determined by finding the proper trade-off between the costs and the benefits of inspection. The amount of inspection is optimal when the total cost of conducting the inspection *and* of passing defectives is minimized, as indicated in Fig. 10.11.[7]

10.8.2 Acceptance Sampling and Its Alternative Plans

Statistical quality control generally uses only the sampling approach to examine the quality of a product. In *acceptance sampling*, the decision whether to accept an entire lot of a product or service is based only on a sample of the lot. By a *lot* we generally mean an amount of material that can be conveniently handled. It may consist of a certain number of items, a case, a day's production, a car load, or such similar quantity. These lots might be described as *convenience lots*. The following two sampling plans are customarily based on convenience lots.

[7] The formula for determining optimal sample size can be found in Sect. 20.4.

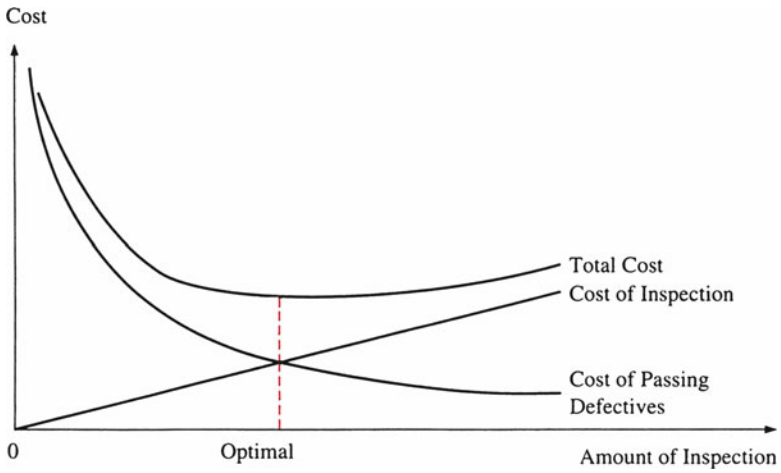


Fig. 10.11 The amount of inspection is optimal when the sum of the cost of inspection and the cost of passing defectives is minimized (*Source*: W. J. Stevenson, *Production/Operations Management*, 3rd ed., 1990, Fig. 16.7, p. 826. Reprinted by permission of Richard D. Irwin)

10.8.2.1 Single-Sampling Plans

In a *single-sampling plan*, one random sample with sample size n is drawn from each lot with N items, and every item in the sample is examined and classified as either good or defective. If any sample contains more than a specified number of defectives, c, then that lot is rejected.

10.8.2.2 Double-Sampling Plans

Double-sampling plans provide for taking a second sample when the results of a first sample are marginal, as is often the case when lots are of borderline quality. Such plans are commonly based on five statistics:

n_1 = size of the first sample
c_1 = acceptance number of defectives for the first sample (n_1)
n_2 = size of the second sample
c_2 = acceptance number of defectives for $n_1 + n_2$
k_1 = retest number for the first sample

For example, say $c_1 = 3$, $c_2 = 8$, $k_1 = 6$, $n_1 = 25$, and $n_2 = 40$. This sample plan dictates the lot size (the size of the initial sample), $n_1 = 25$ items, and it specifies the accept/reject criteria for the initial sample, $c_1 = 3$ and $k_1 = 6$. If 3 or fewer defectives are found, it tells us, accept the lot; if more than 6 defectives are found, reject the lot; and if 4, 5, or 6 defectives are found, take a second sample with sample size $n_2 = 40$.

10.8.2.3 The Advantages of Double-Sampling Plans

A double-sampling plan makes n_1 smaller than the sample size for a single-sampling plan that has essentially the same ability to discriminate between lots of high quality and lots of low quality. This means that good-quality lots of product or service are accepted most of the time on the basis of smaller samples than a comparable single-sampling plan requires. Also, bad lots are generally rejected on the basis of the first sample. A double-sampling plan also gives lots of marginal quality a second chance. This feature appeals to practical-minded production managers.

How well it discriminates between lots of high quality and lots of low quality is an important feature of a sampling plan. The ability of a sampling plan to discriminate can be analyzed and tested, as we will see in Appendix 1 of Chap. 11.

10.8.3 Process Control

Process control is concerned with ensuring that *future* output is acceptable. Toward that end, periodic samples of process output are taken and evaluated. If the output is acceptable, the process is allowed to continue; if the output is not acceptable, the process is stopped and corrective action is instituted. The basic elements of control for quality, costs, labor power, accidents, and just about anything else are the same:

1. Define what is to be controlled.
2. Consider how measurement for control will be accomplished.
3. Define the level of quality that is to be the standard of comparison.
4. Distinguish between random and nonrandom variability and determine what process is out of control.
5. Take corrective action and evaluate that action.

Among these elements of quality control, determining whether an output process is *in control* or *out of control* is the most important task. To do so, we need to analyze the statistical product distribution of the process. If the process variation is due to random variability (common causes of variation), then the process is in control. If the process variation is due to nonrandom variability (special causes of variation), then the process is out of control. The control charts discussed in the next section can help us differentiate between process variation attributable to common causes and variation due to special causes.

10.9 Control Charts for Quality Control

Control charts were first proposed by Walter Stewart at Bell Laboratories in the 1920s. More recently the control chart has become a principal tool in assisting businesses in Japan, the United States, and elsewhere in their quality and productivity efforts.

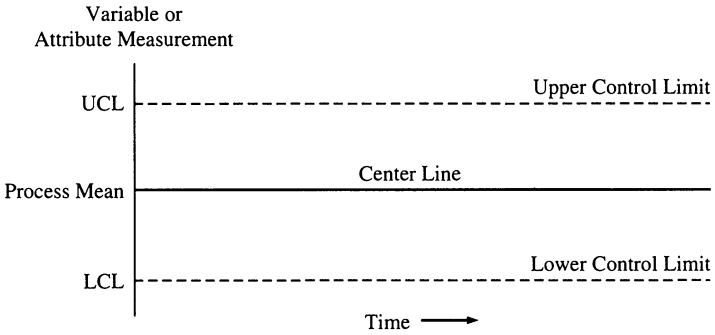


Fig. 10.12 The structure of a control chart (*Source*: Evans and Lindsay (1989), Fig. 12.3, p. 318)

A *control chart* is a graphical tool for describing the state of control of a process. Figure 10.12 illustrates the general structure of a control chart. Time is measured on the horizontal axis, which usually corresponds to the average value of the quality characteristic being measured on the vertical axis. Two other horizontal lines (usually dashed) represent the *upper control limit* (UCL) and the *lower control limit* (LCL). These limits are chosen such that there is a high probability (generally greater than 0.99) that sample values will fall between them if the process is in control. Samples are chosen over time, plotted on the appropriate chart, and analyzed. Basic statistical concepts used to draw the control charts include the expected value, standard deviation, and confidence interval, which we discussed earlier in this chapter. The control charts we will examine in this section are the $\bar{X}$-chart, the $\bar{R} - chart,$ and the S-chart.

10.9.1 $\bar{X}$ -Chart

A statistical quality control chart for means ($\bar{X}$-chart) relies on the interval estimate concept discussed in Sects. 10.4 and 10.5. The $\bar{X}$-chart is used to depict the variation in the centering process. Say copper rods that have a sample mean diameter $\bar{X}$ of 3 cm and a given[8] standard deviation σ_x of .15 cm are being produced by a particular process. It is known that the diameter measurements are normally distributed. The quality control manager might like to determine what control limits will include 99.73 % of the sample mean if the process is generating random output (around the mean) for sample size $n = 36$.

To solve this problem, the quality control department establishes an upper control limit (UCL) and a lower control limit (LCL) in accordance with Eq. 10.7. For the upper control limit,

[8] In quality control, *given standard deviation* means the quality standards of a product are given.

$$P(\bar{X} > \text{UCL}) = P\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} > \frac{\text{UCL} - \mu}{\sigma_{\bar{X}}}\right)$$
$$= P\left(z > \frac{\text{UCL} - 3}{.15/\sqrt{36}}\right) = .00135$$

For the lower control limit,

$$P(\bar{X} < \text{LCL}) = \left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < \frac{\text{LCL} - \mu}{\sigma_{\bar{X}}}\right)$$
$$= \left(z < \frac{\text{LCL} - \mu}{.15/\sqrt{36}}\right) = .00135$$

From Table A3 of Appendix A, we can solve for UCL and LCL as follows[9]:

$$\frac{\text{UCL} - 3}{.025} = 3$$
$$\frac{\text{LCL} - 3}{.025} = -3$$

From these two equations, we obtain

$$\text{UCL} = .075 + 3 = 3.075\,\text{cm}$$
$$\text{LCL} = 3 - .075 = 2.925\,\text{cm}$$

This example shows that we can calculate control limits for the $\bar{X}$ − chart for given standards by using the interval estimate for the population mean μ when σ_X^2 is known, as was discussed in Sect. 10.4.

In quality control, a sequence of k samples with n_j observations each is taken over time on a measurable characteristic of the output of a production process. From this sample, the sample mean $\bar{X}_i$ ($i = 1,2,\ldots,k$) and the overall mean can be defined as

$$\bar{X}_i = \sum_{j=1}^{n} X_{ij}/n \qquad \text{and}$$
$$\bar{X} = \sum_{i=1}^{k} \bar{X}_i/k$$

Taking into account $\bar{X}$, the given standard deviation σ_X, and the logic illustrated in the foregoing example, we see that UCL and LCL can be defined as

$$\text{UCL}_{\bar{X}} = \bar{X} + A\sigma_X \tag{10.14a}$$

[9] In quality control work, control limits are three standard errors on either side of the mean of the sampling distribution. These limits are called 3−σ limits.

$$\text{LCL}_{\bar{X}} = \bar{\bar{X}} - A\sigma_X \tag{10.14b}$$

where $A = 3/\sqrt{n}$, $\bar{\bar{X}}$ = mean of sample means, and σ_X = given process standard deviation. Equations 10.14a and 10.14b can be used to construct the $\bar{X}$-chart when standards are given (σ_X is assumed to be known). The value of A can be found in Table A13 of Appendix A.

If the process standard deviation is not known, then it must be estimated from sample standard deviations as indicated in Eq. 10.15:

$$\bar{s} = \sum_{i=1}^{k} s_i/k \tag{10.15}$$

where

$$s_i = \sqrt{\sum_{j=1}^{n} \frac{(X_{ij} - \bar{X}_j)^2}{n-1}}$$

and X_{ij} is the jth observation in the ith sample. The sample standard deviation, of course, is a biased estimator of population standard deviation. If the population distribution is normal, it can be shown that

$$E(\bar{s}) = C_4\sigma_X \tag{10.16}$$

where C_4 is a number that can be calculated as a function of the sample size n. Then the standard deviation estimate for X can be defined as $\bar{s}/(C_4\sqrt{n})$. This information enables us to write Eqs. 10.14a and 10.14b as[10]

$$\text{UCL}_{\bar{X}} = \bar{\bar{X}} + A_3\bar{s} \tag{10.17a}$$

$$\text{LCL}_{\bar{X}} = \bar{\bar{X}} - A_3\bar{s} \tag{10.17b}$$

where $A_3 = 3\bar{s}/(C_4\sqrt{n})$, which can be found in Table A13.

In quality control, we often use sample range instead of sample standard deviation to estimate both upper and lower control limits for our $\bar{X}$ chart. Recall from Eq. 4.11 in Chap. 4 that the range R of a sample is the difference between the maximum and

[10] If the underlying sampling is the Poisson distribution as discussed in Sect. 6.7, then the $\bar{x}$ and $\bar{s}$ can be defined as $\bar{c}$ and $\sqrt{\bar{c}}$, respectively ($\bar{c}$ is defined as the mean number of defects per unit). In this situation the $\bar{X}$ -chart defined in Eqs. 10.17a and 10.17b is called the C-chart (see Evans and Lindsay, 1989, pp. 366–368).

minimum measurement in the sample. The range can be used to obtain an unbiased estimator for σ_X defined as follows[11]:

$$\hat{\sigma}_X = \frac{\overline{\mathrm{RG}}}{d_2}$$

where

$$\overline{\mathrm{RG}} = \sum_{i=1}^{k} R_i/k, \qquad R_i$$

equals the range in ith sample and d_2 is a constant that can be found in Table A13.

If sample range instead of sample standard deviation is used to replace the process standard deviation, then the control limits can be defined as

$$\mathrm{UCL}_{\bar{X}} = \bar{X} + A_2\overline{\mathrm{RG}} \tag{10.18a}$$

$$\mathrm{LCL}_{\bar{X}} = \bar{X} - A_2\overline{\mathrm{RG}} \tag{10.18b}$$

where $A_2 = 3/d_2\sqrt{n}$ can be found in Table Al3 of Appendix A and d_2 is a function of sample size n.

10.9.2 $\bar{R}$ -Chart and S-*Chart*

Besides the variation of centering (mean), we are also interested in the variation of dispersion (standard deviation or range) in quality control. The $\bar{R}$ *-chart* is used to depict the variation of the ranges of the samples. The *S-chart* is used to depict the variation of standard deviation. In other words, both the *S*-chart and $\bar{R}$ -chart can be used to detect changes in process variation. The $\bar{R}$ -chart is used more frequently than the *S*-chart because the range is much easier to calculate than the standard deviation.

It can be shown that $E(\bar{s}) = C_4\sigma_X$, as we have noted, and that the standard deviation of $\bar{s}$ is $\sigma_X\sqrt{1 - C_4^2}$. When no standards are given, we use $\bar{s}$ as an estimate of $C_4\sigma_X$. Then the upper and lower limits of the *S*-chart can be denned as

$$\mathrm{UCL}_s = B_4\bar{s} \tag{10.19a}$$

$$\mathrm{LCL}_s = B_3\bar{s} \tag{10.19b}$$

[11] See T. T. Ryan (1989), *Statistical Methods for Quality Improvement* (New York: Wiley), for a detailed discussion of this relationship.

where $B_4 = 1 + 3\sqrt{1 - C_4^2}/C_4$ and $B_3 = 1 - 3\sqrt{1 - C_4^2}/C_4$. Both can be found in Table A13.

If no standards are given, the upper and lower limits of the $\bar{R}$ -chart can be defined as

$$\text{UCL}_{\bar{R}} = D_4\overline{\text{RG}} \tag{10.20a}$$

$$\text{LCL}_{\bar{R}} = D_3\overline{\text{RG}} \tag{10.20b}$$

where $\overline{\text{RG}}$ is the average of sample ranges and where $D_4 = 1 + 3d_3/d_2$ and $D_3 = 1 - 3d_3/d_2$ can be found in Table A13.

The $\bar{X}$ -chart, S-chart, and $\bar{R}$ -chart, then, all use the confidence interval concept to construct upper and lower limits. Now we will use quality control data on Consolidated Auto Supply Company to show how these control charts are constructed.

Application 10.1 $\bar{X}$ -Chart, $\bar{R}$ -Chart, and S-Chart for Consolidated Auto Supply Company. The quality control manager has measured the size of U-bolts by taking samples of 5 every hour over 3 shifts.[12] The sample is presented in Table 10.3, which also shows the mean and range of each sample. How do we perform this statistical quality control analysis?

To construct our $\bar{X}$ -chart and $\bar{R}$ -chart, we first compute the average mean $\bar{X}$ and average range $\overline{\text{RG}}$ as follows:

$$\bar{X} = \frac{10.7 + 10.77 + \cdots + 10.66}{24} = 10.7171$$

$$\overline{\text{RG}} = \frac{.20 + .20 + \cdots + .10}{24} = .1792$$

Using the information on $\bar{X}$, $\overline{\text{RG}}$, and $n = 5$, we calculate control limits for our $\bar{X}$ -chart and $\bar{R}$ -chart in accordance with Eqs. 10.18a, 10.18b, 10.20a, and 10.20b.

$$\begin{aligned}
\text{UCL}_{\bar{X}} &= 10.7171 + .58(.1792) = 10.8210 \\
\text{LCL}_{\bar{X}} &= 10.7171 - .58(.1792) = 10.6132 \\
\text{UCL}_{\bar{R}} &= 2.11(.1792) = .3782 \\
\text{LCL}_{\bar{R}} &= 0(.1792) = 0
\end{aligned}$$

The $\bar{X}$ -chart and $\bar{R}$ -chart for Consolidated Auto Supply Company are displayed in Figs. 10.13 and 10.14, respectively. These control charts can be used to do statistical quality control analysis.

[12] This example is drawn from J. R. Evans and W. M. Lindsay (1989). *The Management and Control of Quality* (St. Paul, MN: West), pp. 317–323 and pp. 359–360.

Table 10.3 Sample means and ranges for Consolidated Auto Supply Company

Sample	Observations					Mean	Range
1	10.65	10.70	10.65	10.65	10.85	10.70	0.20
2	10.75	10.85	10.75	10.85	10.65	10.77	0.20
3	10.75	10.80	10.80	10.70	10.75	10.76	0.10
4	10.60	10.70	10.70	10.75	10.65	10.68	0.15
5	10.70	10.75	10.65	10.85	10.80	10.75	0.20
6	10.60	10.75	10.75	10.85	10.70	10.73	0.25
7	10.60	10.80	10.70	10.75	10.75	10.72	0.20
8	10.75	10.80	10.65	10.75	10.70	10.73	0.15
9	10.65	10.80	10.85	10.85	10.75	10.78	0.20
10	10.60	10.70	10.60	10.80	10.65	10.67	0.20
11	10.80	10.75	10.90	10.50	10.85	10.76	0.40
12	10.85	10.75	10.85	10.65	10.70	10.76	0.20
13	10.70	10.70	10.75	10.75	10.70	10.72	0.05
14	10.65	10.70	10.85	10.75	10.60	10.71	0.25
15	10.75	10.80	10.75	10.80	10.65	10.75	0.15
16	10.90	10.80	10.80	10.75	10.85	10.82	0.15
17	10.75	10.70	10.85	10.70	10.80	10.76	0.15
18	10.75	10.70	10.60	10.70	10.60	10.67	0.15
19	10.65	10.65	10.85	10.65	10.70	10.70	0.20
20	10.60	10.60	10.65	10.55	10.65	10.61	0.10
21	10.50	10.55	10.65	10.80	10.80	10.66	0.30
22	10.80	10.65	10.75	10.65	10.65	10.70	0.15
23	10.65	10.60	10.65	10.60	10.70	10.64	0.10
24	10.65	10.70	10.70	10.60	10.65	10.66	0.10

Source: Evans and Lindsay (1989), Table 12.2, p. 319

The location of points and patterns of points in a control chart makes it possible to determine, with only a small chance of error, whether a process is in a state of statistical control. In both Figs. 10.13 and 10.14, the chance that a sample mean or range will fall outside the control limits is only .27 %. Therefore, the first indication that a process may be out of control is a point lying outside the control limits. In the $\bar{R}$ -chart, sample 11 is outside the UCL limit, indicating that the variability of the process has changed. In this case, it is found that the change in process variability is due to the fact that a substitute operator was used. In the $\bar{X}$ -chart, sample 21 is outside the LCL, and samples 18 through 24 are all on one side. This indicates that the process mean has shifted. In this case, it is found that the shift in process mean occurred because nonconforming material was used.

From Table 10.3, we find that the standard deviation of the observation is $\bar{s} = .07958$. Substituting $\bar{X} = 10.7171$, $\bar{s} = .07958$, $n = 5$, $A_3 = 1.427$, $B_3 = 0$, and $B_4 = 2.089$ (from Table A13 in Appendix A) into Eqs. 10.19a, 10.19b, 10.17a, and 10.17b, we obtain the following control limits for the S-chart and $\bar{X}$ -chart:

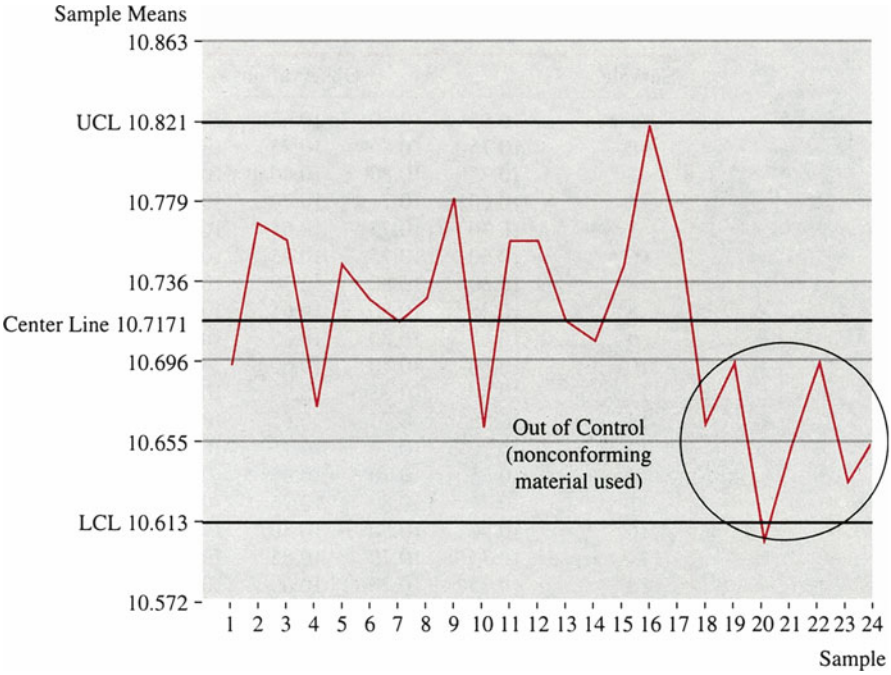


Fig. 10.13 $\bar{X}$-chart for Consolidated Auto Supply Company (*Source*: Evans and Lindsay (1989), Fig. 12.5, p. 321)

$$\begin{aligned}
\text{UCL}_s &= 2.089(.07958) = .1662 \\
\text{LCL}_s &= 0(.07958) = 0 \\
\text{UCL}_{\bar{X}} &= 10.7171 + 1.427(.07958) = 10.8307 \\
\text{LCL}_{\bar{X}} &= 10.7171 - 1.427(.07958) = 10.6035
\end{aligned}$$

The $\bar{X}$-chart and S-chart are displayed in Figs. 10.15 and 10.16, respectively. Both charts indicate that the product process is in a state of statistical control.

Application 10.2 Establishing Statistical Control and Determining Process Capability. Control charts have three basic applications: (a) to establish a state of statistical control, (b) to determine process capability, and (c) as a monitoring device to signal the existence of assignable causes in order to maintain a state of statistical control. The process capability represents the natural variation resulting from using a given combination of people, machinery, materials, methods, and management.

Both Figs. 10.13 and 10.14 have indicated that the process is out of control, and this has been discussed in detail in Application 10.1. These results imply that the control limits will be in error. To revise the control limits for $\bar{X}$ -chart, we exclude points 18, 19, . . ., and 24. The new control limits of $\bar{X}$-chart are established as

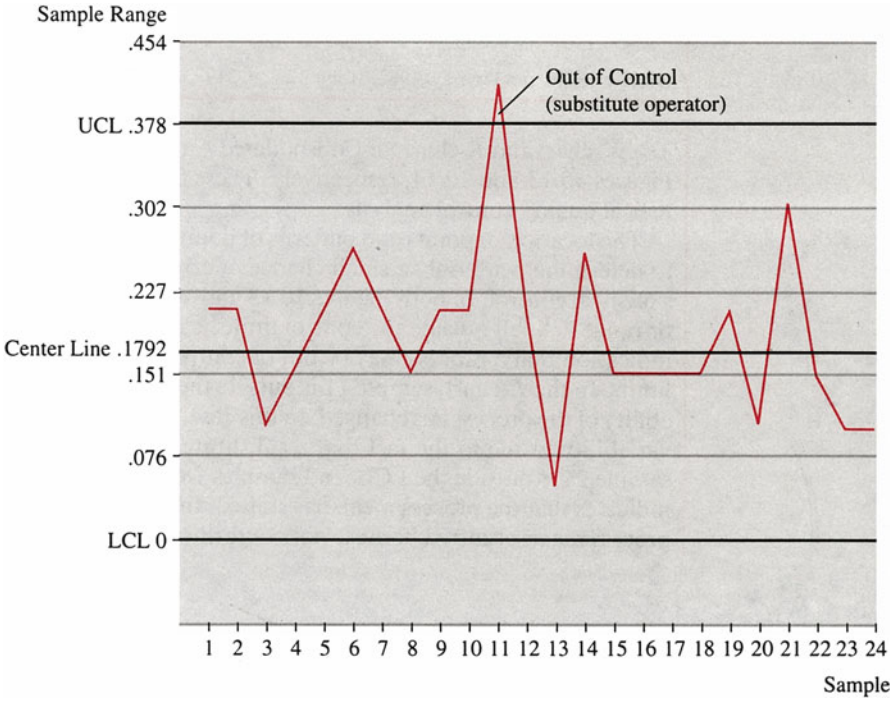


Fig. 10.14 $\bar{R}$-chart for Consolidated Auto Supply Company (*Source*: Evans and Lindsay (1989), Fig. 12.4, p. 320)

$$UCL_{\bar{X}} = 10.7394 + .58(.1696) = 10.8378$$
$$LCL_{\bar{X}} = 10.7394 - .58(.1696) = 10.6410$$

From the revised $\bar{X}$-chart established from these new control limits, it can be shown that the control process is in control.

To revise the control limits for $\bar{R}$-chart, we exclude point 11. It can be shown that $\bar{R}$ is now equal to .1696. The new control limits for the range are

$$UCL_{\bar{R}} = 2.11(.1696) = .3579$$
$$LCL_{\bar{R}} = 0(.1696) = 0$$

After a process has been brought to a state of statistical control by eliminating special causes of variation, we can determine the capability of the process. This is a simple calculation based on the average range. However, a critical assumption is that the distribution of process output follows a normal probability distribution; otherwise, we cannot invoke the central limit theorem as discussed in Sect. 8.6 in Chap. 8. If this is not the case, the results of this calculation will not be correct, and

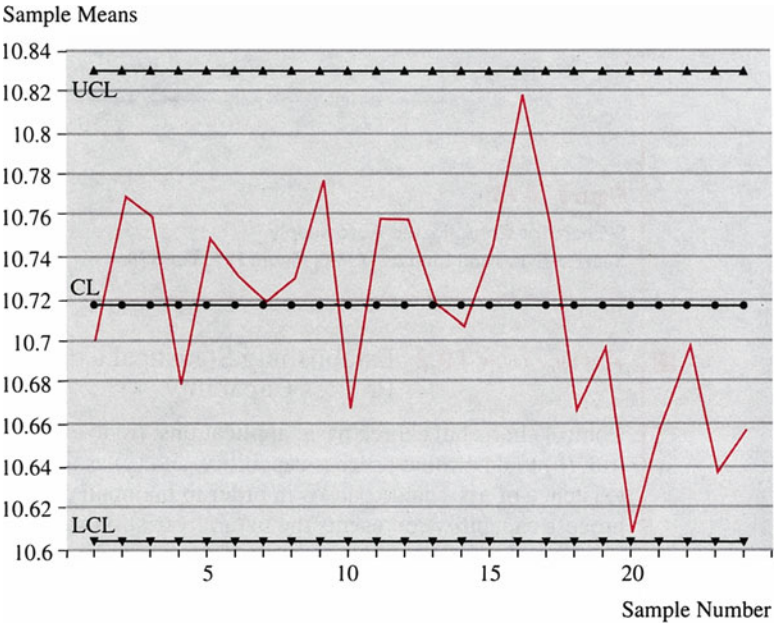


Fig. 10.15 $\bar{X}$ – *chart* in terms of $\bar{s}$ for Consolidated Auto Supply Company (*Source*: Evans and Lindsay (1989), Fig. 12A.2, p. 377)

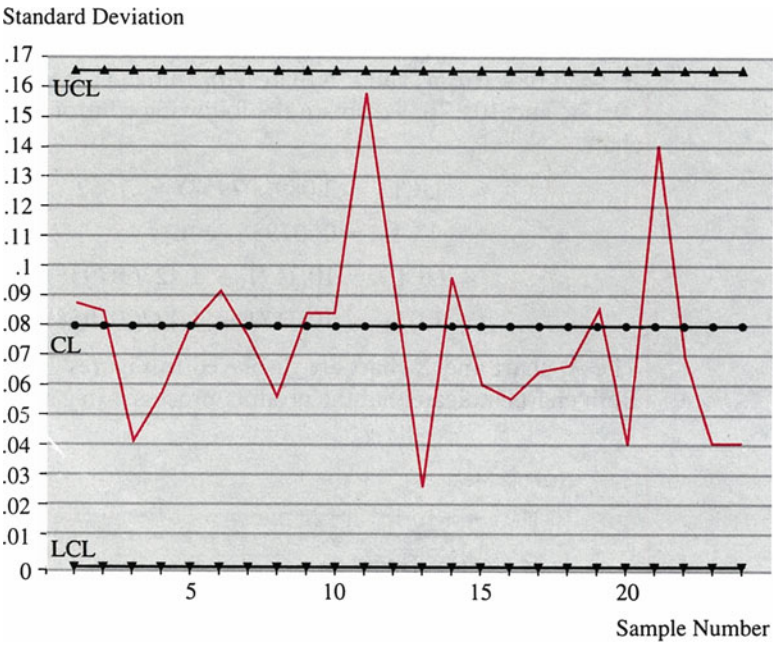


Fig. 10.16 S-chart for Consolidated Auto Supply (*Source*: Evans and Lindsay (1989), Fig. 12A.3, p. 378)

different techniques must be used. Under the normality assumption, the process standard deviation σ can be estimated as follows:

$$\sigma = \overline{RG}/d_2$$

where values for d_2 depend on the sample size and are also listed in Table A13. For $n = 5$ and $d_2 = 2.326$, we have

$$\sigma = .1696/2.326 = .0729$$

Three-sigma limits on the natural process variation are given by $\bar{X} \pm 3\sigma$ or 10.7394–3(.0729) = 10.5207 and 10.7394 + 3(.0729) = 10.9581. Using these new limits, we can compute the percentage of nonconforming parts if the specifications are 10.55–10.90. We leave it to you to verify that the *Z*-values corresponding to 10.55and 10.90 are 2.60 and 2.21, respectively. Using Table A3, we find that the areas to the left and right of these values under the standard normal density are 0.0119 and 0.0136. Therefore, the proportion of nonconforming U-bolts that are expected to be produced by this process is 0.0119 + 0.0136 = 0.0255, or 2.55 %. This cannot be improved unless the design standards are changed or the process is improved.

There is one word of caution which we wish to emphasize. Control limits are often confused with specification limits. Specification dimensions are usually stated in relation to individual parts for "hard" goods, such as automotive hardware. However, in other applications, such as in chemical processes, specifications are stated in terms of average characteristics. Thus, control charts might mislead one into thinking that if all sample averages fall within the control limits, all output will be conforming. This is not true. Control limits relate to *averages*, while specification limits relate to individual measurements. It is possible that a sample average falls within the upper and lower control limits and yet some of the individual observations are out of specification. Since $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, control limits are narrower than the natural variation of the process and do not represent process capability.

10.9.3 Control Charts for Proportions

Control charts for proportions are used when the process characteristic is counted rather than measured. The *P-chart* is used to measure the percentage of defectives generated by a process. The theoretical basis for a *P*-chart is the binomial distribution (see Chap. 6). Conceptually, a *P*-chart is constructed and used in much the same way an $\bar{X}$ -chart is.

Let $\hat{p}_i$ be the fraction of defectives in the *i*th sample with *n* observations; then the center line on a *P*-chart is the average fraction of defectives for *k* samples as defined as:

Table 10.4 Sorting errors at the Newton Branch Post Office

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Errors	3	1	0	0	2	5	3	6	1	4	0	2	1	3	4
Sample	16	17	18	19	20	21	22	23	24	25					
Errors	1	1	2	5	2	3	4	1	0	1					

Source: Evans and Lindsay (1989), Table 12.4, p. 333

$$\bar{P} = \frac{\sum_{i=1}^{k} \hat{P}_i}{k}$$

The standard deviation associated with $\bar{P}$ is

$$s_{\bar{P}} = \sqrt{\bar{P}(1-\bar{P})/n}$$

and the upper and lower control limits are

$$\text{UCL}_{\bar{P}} = \bar{P} + 3s_{\bar{P}} \tag{10.21a}$$

$$\text{LCL}_{\bar{P}} = \bar{P} - 3s_{\bar{P}} \tag{10.21b}$$

Application 10.3 *P*-Chart for Quality Control at the Newton Branch Post Office. In the post office, operators use automated sorting machines that read the ZIP code on a letter and divert the letter to the proper carrier route.[13] Over 1 month's time, 25 samples of 100 letters were chosen, and the numbers of errors were recorded. This information is summarized in Table 10.4. The fraction nonconforming is found by dividing the number of errors by 100. The average fraction nonconforming, $\bar{P}$, is determined to be

$$\bar{P} = \frac{.03 + .01 + \cdots + .01}{25} = .022$$

The standard deviation is

$$s_{\bar{P}} = \sqrt{\frac{.022(1-.022)}{100}} = .01467$$

Thus, the upper control limit, UCL_p, is .022 + 3(.01467) = .066, and the lower control limit, $LCP_{\bar{p}}$, is .022–3(.1467) = −.022. Because this latter figure is negative, zero is used instead. The control chart for the Newton Branch Post Office is

[13] This example is drawn from Evans and Lindsay (1989), pp. 332–333.

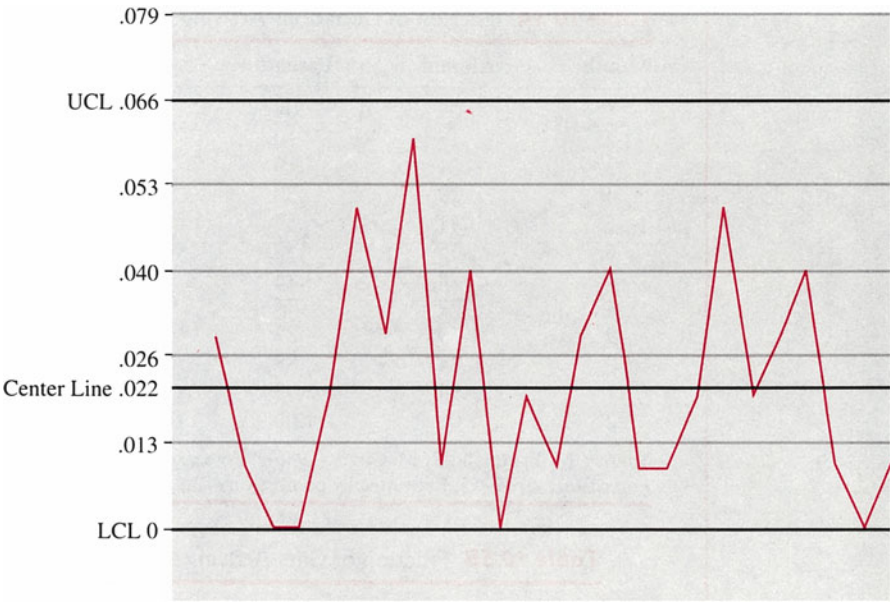


Fig. 10.17 *P*-chart for Newton Branch Post Office (*Source*: Evans and Lindsay (1989), Fig. 12.14, p. 333)

shown in Fig. 10.17. The sorting process appears to be in control. If any values had been found above the upper control limit or if an upward trend were evident, it might indicate a need for operators with more experience or for more training of the operators.

10.10 Further Applications

In the last section we showed how interval estimates for the mean, proportion, and standard deviation can be used to construct quality control charts. The next two examples show how interval estimates can be used for other business applications.

Application 10.4 Using Interval Estimates to Evaluate Donors and Donations Models. Britto and Oliver (1986) developed models to forecast (1) the total numbers of donors, gifts, and donations by the end of each year and (2) the cumulative numbers of donors, gifts, and donations received up to and including the *t*th month for the Berkeley Engineering Fund.[14]

[14] M. Britto and R. M. Oliver (1986), "Forecasting Donors and Donations," *Journal of Forecasting* 5, 39–55.

Table 10.5A Fraction of donations arriving in or before month *t*

Month	Alumni	Parents	Faculty	Friends
1	0.09	0.04	0.01	0.05
2	0.15	0.10	0.01	0.12
3	0.21	0.19	0.04	0.20
4	0.29	0.30	0.16	0.28
5	0.41	0.37	0.28	0.37
6	0.65	0.58	0.77	0.56
7	0.74	0.75	0.90	0.71
8	0.77	0.77	0.94	0.72
9	0.79	0.87	0.96	0.73
10	0.84	0.88	0.97	0.75
11	0.94	0.93	0.98	0.95
12	1.00	1.00	1.00	1.00

Source: M. Britto and R. M. Oliver (1986), "Forecasting Donors and Donations." *Journal of Forecasting* 5, 39–55. Reprinted by permission of John Wiley & Sons, Ltd

Table 10.5B Fraction of gifts arriving in or before month *t*

Month	Alumni	Parents	Faculty	Friends
1	0.08	0.04	0.01	0.05
2	0.13	0.10	0.01	0.12
3	0.19	0.18	0.04	0.20
4	0.26	0.28	0.16	0.29
5	0.38	0.35	0.27	0.39
6	0.61	0.56	0.75	0.68
7	0.70	0.73	0.89	0.76
8	0.72	0.76	0.93	0.78
9	0.76	0.86	0.95	0.78
10	0.80	0.87	0.95	0.80
11	0.90	0.90	0.96	0.95
12	1.00	1.00	1.00	1.00

Source: M. Britto and R. M. Oliver (1986), "Forecasting Donors and Donations," *Journal of Forecasting*, *5*, 39–55. Reprinted by permission of John Wiley & Sons, Ltd

Forecasts are based on data from previous campaigns because identical mailings were used from 1982 to 1984. Monthly proportions of total giving have been stable from year to year, as shown in Tables 10.5A and 10.5B. For each mailing date, the forecasters determined the distribution for the number of gifts for each of the four subgroups (see Table 10.6), as well as estimates of the mean and standard deviation of gift size (see Table 10.7).

Parent data from 1982 to 1983 and 1983 to 1984 were used to test whether the Poisson distribution on which the model is based is acceptable. Using both Poisson tables and a normal approximation (discussed in Chap. 7), Britto and Oliver constructed 95 % confidence intervals as shown in Figs. 10.18 and 10.19. For

Table 10.6 Prior expected numbers of donations and gifts in 1984–1985

	Donors	Gifts
Alumni	2,807	3,265
Parents	248	277
Faculty	117	129
Friends	87	93

Source: M. Britto and R. M. Oliver (1986), "Forecasting Donors and Donations," *Journal of Forecasting*, 5, 39–55. Reprinted by permission of John Wiley & Sons, Ltd

Table 10.7 1983–1984 Mean and standard deviation of gift size

	Mean	Standard deviation
Alumni	215	1,820
Parents	200	930
Faculty	225	445
Friends	505	1,215

Source: M. Britto and R. M. Oliver (1986). "Forecasting Donors and Donations," *Journal of Forecasting*, 5, 39–55. Reprinted by permission of John Wiley & Sons, Ltd

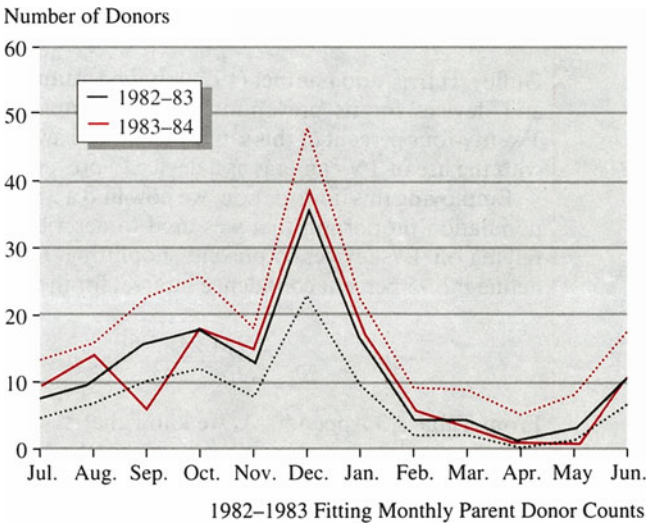


Fig. 10.18 1982–1983 Fitting parent donor counts

both years, the actual donor counts for all months except September fell within 95 % confidence limits. This leads us to believe that the Poisson distribution assumption is a good one.

Application 10.5 Shoppers' Attitudes Toward Shoplifting and Shoplifting Prevention Devices. Guffey, Harris, and Laumer (1979) studied attitudes of

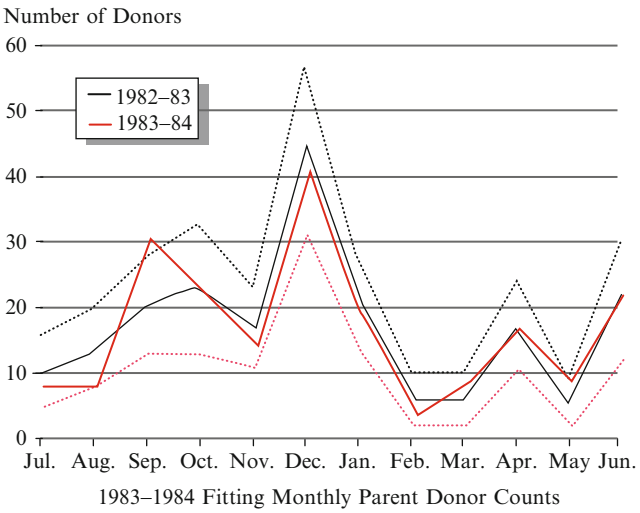


Fig. 10.19 1983–1984 Fitting parent donor counts

shoppers toward shoplifting and devices for its prevention.[15] They sampled 403 shopping center patrons. Twenty-four percent of this sample expressed awareness of and uncomfortableness with the use of TV cameras as a device to prevent shoplifting.

Employing this information, we now find a 95 % confidence interval of the population proportion that was used to describe the attitudes of shoppers about relaying on TV devices to prevent shoplifting. Following Eq. 10.11, we can define the 95 % confidence interval for the true proportion p as

$$.24 - z_{.025}\sqrt{\frac{(.24)(.76)}{403}} < z_{.025}\sqrt{\frac{(.24)(.76)}{403}}$$

From Table A3 Appendix A, we know that $z_{.025} = 1.96$. Substituting $z_{.025} = 1.96$ onto the previous equation, we obtain the 95 % confidence interval for p as

$$.240 - (1.96)(.021) < p < .240 + (1.96)(.021)$$

or

$$.199 < p < .281$$

Hence, the interval for p ranges from 19.9 % to 28.1 %.

[15] H. L. Guffey, J. R. Harris, and J. F. Laumer (1979), "Shopper Attitudes Toward Shoplifting and Shoplifting Prevention Devices," *Journal of Retailing* 55, 75–99.

10.11 Summary

In this chapter, we used concepts discussed in Chaps. 7, 8, and 9 to show how point estimates and confidence intervals are constructed. Applications of point estimation and of the confidence interval in constructing quality control charts and other business applications were also discussed. In Chap. 11, we will draw on the concepts discussed in this chapter to test hypotheses about sample point estimates.

Questions and Problems

1. For the following results from samples drawn from normal populations, what are the best estimates for the mean, the variance, the standard deviation, and the standard deviation of the mean?

   (a) $n = 9, \Sigma X_i = 36, \Sigma\ (X_i - \bar{X})^2 = 288$
   (b) $n = 16, \Sigma\ X_i = 64, \Sigma\ (X_i - \bar{X})^2 = 180$
   (c) $n = 25, \Sigma\ X_i = 500, \Sigma X_i^2 = 12{,}400$

2. For each of the following samples drawn from normal populations, find the best estimates for μ, σ^1, σ, and the standard deviation of $\bar{X}$.

   (a) 4, 10, 2, 8, 4, 14, 12, 8, 30
   (b) –4, 2, –6, 0, 6, 2, 4, 0, –4
   (c) 6, 15, 13, 21, 10, 17, 12

3. Find the value of $z_{\alpha l2}$ for the following values of α.

   (a) $\alpha = .01$ *b. c.* $\alpha = .002$
   (b) $\alpha = .03$
   (c) $\alpha = .002$

4. A stockbroker has taken a random sample of 4 stocks from a large population of low-priced stocks. Stock prices for this population are normally distributed. The sample prices of the 4 stocks are \$5, \$12, \$17, and \$10.

   (a) Calculate a point estimate of the population mean.
   (b) Calculate a point estimate of the population variance. What is your estimate for a population standard deviation?
   (c) Calculate a point estimate of the proportion of stocks in this population that are priced at \$10 or more.

5. A 90 % confidence interval for the population mean time (in minutes) needed to finish a certain assembly process is $90 < \mu_x < 130$.

   (a) Sketch this interval, indicating the margin for sampling error.
   (b) If the sample size was $n = 25$, what was the sample standard deviation?
   (c) To interpret this confidence interval, what did you have to assume about the population? Why?

6. Assuming you have samples from normal populations with known variance, find
   (a) The degree of confidence used if $n = 19$, $\sigma = 8$, and the total width of a confidence interval for the mean is 3.29 units
   (b) The sample size when $\alpha^2 = 100$ and the 95 % confidence interval for the mean is from 17.2 to 22.8
   (c) The known variance when $n = 100$ and the 98 % confidence interval for the mean is 28.26 units in width
7. (a) Find the value of t such that the probability of a larger value is .005 when the value for the degrees of freedom is very large.
   (b) Find the value of t such that the probability of a smaller value is .975 when the value for the degrees of freedom is very large (infinite).
   (c) Are the t values essentially the same as corresponding Z-values when the value of the degrees of freedom is very large?
8. A poll reported that 48 % of probable voters seem determined to vote against the president. Assume that this sample was based on a random selection of 789 probable voters. Construct a 99 % confidence interval for the probable voters who seem determined to vote against the president.
9. A survey of low-income families in New Jersey was designed to determine the average heating costs for a family of 4 during January and February. Heating costs are known to have a standard deviation of $25.75. The economists conducting the study wish to construct a 95 % confidence interval with a margin for sampling error of no more than $3.95. Find the appropriate sample size.
10. A company has just installed a new automatic milling machine. The time it takes the machine to mill a particular part is recorded for a sample of 9 observations. The mean time is found to be $\bar{X} = 8.50$, and $S^2 = .0064$. Find a 90 % confidence interval for the unknown mean time for milling this part.
11. A survey indicated that companies with fewer than 1,000 employees are expected to increase their spending by 20.4 %. Form a 99 % confidence interval for the unknown mean increase, assuming that the sample standard deviation is 6.8 % and the sample size is 346.
12. A study conducted in 1984 reported that the median pay in the United States was $18,700. What difficulties do you see in using this type of study for assessing incomes? Would you be willing to use $18,700 as a point estimate of the central location of US incomes?
13. What is a point estimate? What is a point estimator? What is point estimation? How are these concepts related to the concepts of sampling that we discussed in Chap. 9?
14. What is an unbiased estimator? What is an efficient estimator? What is a consistent estimator? Why are these concepts important?
15. Briefly explain why we sometimes construct confidence intervals for the population mean.

16. Explain what happens to the size of the confidence interval when
    (a) The standard deviation increases.
    (b) The standard deviation decreases.
    (c) The probability content $(1-\alpha)$ increases from 95 % to 99 %.
    (d) The sample size increases from 100 to 1,000.
17. Labor economists at the Department of Labor say they have 95 % confidence that factory workers' earnings will lie between $22,000 and $61,000. Explain what this means.
18. A real estate agent in Connecticut is interested in the mean home price in the state. A random sample of 50 homes shows a mean home price of $175,622 and a sample standard deviation of $37,221. Construct a 95 % confidence interval for the mean home price.
19. Reconstruct the confidence interval for the mean home prices given in question 18, but this time construct a 99 % confidence interval. What happens to the size of the confidence interval?
20. Again, use the information given in question 18. This time assume that 100 homes are randomly sampled instead of 50. Construct a 95 % confidence interval for the mean home price. What happens to the size of the confidence interval?
21. Again, use the information given in question 18. This time, assume that the sample standard deviation is $28,000. Construct a 95 % confidence interval for the mean home price. What happens to the size of the confidence interval?
22. An auditor randomly samples 75 accounts receivable of a company and finds a sample mean of $128 with a sample standard deviation of $27. Construct a 90 % confidence interval for the mean accounts receivable.
23. A random sample of 300 residents of a town shows that 55 % believe the mayor is doing a good job. Construct a 95 % confidence interval for the proportion of all residents who believe the mayor is doing a good job.
24. A random sample of 200 students at Academic University finds the sample mean grade point average to be 3.10 with a standard deviation of 0.80. Construct a 99 % confidence interval for the mean grade point average.
25. An insurance company is interested in the average claim on its auto insurance policies. It believes the claims are normally distributed. Using the last 37 claims, it finds the mean claim to be $1,270 with a standard deviation of $421. Construct a 95 % confidence interval for the mean claim on all policies.
26. A random sample of the luggage of 30 passengers of Fly Me Airlines finds that the mean weight of the luggage is 47 lb with a standard deviation of 8 lb. Construct a 90 % confidence interval for the mean weight of Fly Me Airlines luggage.
27. A bank manager finds from reviewing her records that the amount of money deposited on Saturday morning is normally distributed with a standard deviation of $150. A random sample of 7 customers reveals the following amounts deposited on Saturday morning:
    $825 $972 $311 $1,212 $150 $1,800 $725

(a) Find a 95 % confidence interval for the mean amount of deposits by using the MINITAB program.
(b) Find a 90 % confidence interval for the mean amount of deposits by using MINITAB gain. Compare your answer to the confidence interval you computed in part (a). Which is larger?

28. Redo question 27, parts (a) and (b), this time assuming the population standard deviation is unknown. Use MINITAB.
29. A quality control engineer believes that the life of light bulbs for his company is normally distributed with a standard deviation of 100 h. A random sample of 10 light bulbs gives the following information on the life of the light bulbs:
1.000 h; 1,200 h; 600 h; 400 h; 900 h; 500 h; 1,520 h; 1,800 h; 300 h; 525 h

(a) Find a 90 % confidence interval for the mean life of the light bulbs.
(b) Suppose the standard deviation is not known. Construct a 90 % confidence interval for the mean life of the light bulbs.

30. Managers at the Smooth Ride Car Rental Company are interested in the mean number of miles that people drive per day. From past experience, they know that the standard deviation is 75 miles. A random sample of 6 car rentals shows that the people drove the following numbers of miles: 152, 222, 300, 84, 90, and 122. Construct a 99 % confidence interval for the mean number of miles driven.
31. A credit manager at the Bargain Basement Department Store is interested in the proportion of customers who pay their credit card balances in full each month. A random sample of 200 customers indicates that 95 paid their balance in full each month. Construct a 99 % confidence interval for the proportion of customers who pay their balances in full each month.
32. Construct point estimates for the following situations:

(a) A labor union randomly samples 75 of its members and finds that 40 favor the new contract. Estimate the proportion of all workers who favor the new contract.
(b) An economics professor randomly samples 100 students in her class and finds that 70 do not know the meaning of *elasticity*. Estimate the proportion of all students in her class who cannot define this term.

33. An auditor randomly samples 50 accounts payable of a company and finds a sample mean of $1,100 with a sample standard deviation of $287. Construct a 90 % confidence interval for the mean accounts payable.
34. A random sample of 250 residents of a town shows that 55 % favor a bond issue to finance new school construction. Construct a 99 % confidence interval for the proportion of all residents who favor the bond issue.
35. A random sample of 500 students at Average College finds the sample mean combined-SAT score to be 1.050 with a standard deviation of 120. Construct a 90 % confidence interval for the mean SAT score.

36. Reviewing his records, a grocery store manager finds that the amount of money spent shopping on Friday evenings is normally distributed with a standard deviation of $22. A random sample of 5 customers reveals the following amounts spent shopping on Friday night: $125, $72, $15, $88, and $96.
    (a) Find a 95 % confidence interval for the mean amount of money spent shopping.
    (b) Find a 90 % confidence interval for the mean amount of money spent shopping. Compare your answer to the confidence interval you computed in part (a). Which is larger?
37. A random sample of 75 observations from a population yielded the following summary statistics:

$$\sum x = 1,270 \quad \sum x^2 = 21,520$$

    Construct a 95 % confidence interval for the population mean μ.
38. A random sample of 100 observations from a population yielded the following summary statistics:

$$\sum x = 375 \quad \sum (x_1 - \bar{x})^2 = 972$$

    Construct a 99 % confidence interval for the population mean μ.
39. Suppose a random sample of 40 professional golfers is taken and the mean scoring average of the sample is found to be 72.8 strokes per round with a standard deviation of 1.2 strokes per round. Construct a 99 % confidence interval for the population's mean strokes per round.
40. Suppose a random sample of 10 professional golfers is taken and the mean scoring average of the sample is found to be 71.8 strokes per round with a standard deviation of 1.3 strokes per round. Construct a 90 % confidence interval for the population mean strokes per round.
41. Reconsider the information given in question 40. Suppose now that the population standard deviation is known to be 1.3 strokes per round. Construct a 90 % confidence interval for the population mean strokes per round. Compare your answer to your answer in question 40. Why are they different?
42. Suppose you construct a 95 % confidence interval for the mean of an infinite population. Will the interval always be narrower when σ is known than when σ is unknown?
43. A random sample of 75 observations reveals that the sample mean is 20. You know that the population standard deviation is 5. Construct a 90 % confidence interval for the population mean.
44. In a national survey, 200 cola drinkers were asked to compare Yum Yum Cola to Yuk Yuk Cola. Of the 200 people sampled, 120 preferred Yum Yum. Construct a 95 % confidence interval for the actual proportion of consumers who prefer Yum Yum Cola.

45. The 80 members of a random sample of graduates of Mary's Typing School indicate that their mean salary is $22,500 with a sample standard deviation of $3,100. Construct a 99 % confidence interval for the true mean salary.
46. Suppose a random sample of 30 college students reveals that the mean amount of money spent on textbooks each semester is $ 145 with a standard deviation of $25. Construct a 90 % confidence interval for the mean amount of money that students spend on textbooks each semester.
47. The Better Health Cereal Company produces Healthy Oats cereal. A sample of 100 boxes of this cereal indicates that the mean weight of a box of cereal is 24 oz with a standard deviation of 1 oz. Construct a 99 % confidence interval for the population's mean weight.
48. The Better Health Cereal Company produces Healthy Oats cereal. A sample of 15 boxes of this cereal indicates that the mean weight of a box of cereal is 24 oz with a standard deviation of 1 oz. Construct a 99 % confidence interval for the population mean weight. Compare your answer to your answer in question 47. Why are they different?
49. A sample of 100 former basketball players from Slam Dunk University shows that 55 of the players graduated in 4 years. Construct a 90 % confidence interval for the proportion of basketball players graduating in 4 years from Slam Dunk U.
50. A sample of 20 cups of coffee from a coffee machine has a mean amount of coffee of 6 oz. The standard deviation is known to be .5 oz. Construct a 99 % confidence interval for the mean amount of coffee per cup.
51. Reconsider question 50. This time, assume that the standard deviation is not known and that .5 oz is the sample standard deviation. Again construct a 99 % confidence interval for the mean amount of coffee per cup. Compare your answer to your answer in question 50.
52. Suppose a sample of 500 companies listed on the NYSE is found to contain 327 companies paying dividends that have increased over the last year. Construct a 95 % confidence interval for the mean proportion of companies that paid dividends that increased over the last year.
53. A sample of 100 steel-belted radial tires yields a mean life of 35,000 miles with a sample standard deviation of 4,000 miles. Construct a 90 % confidence interval for the mean life of steel-belted radial tires.
54. Suppose a bowler takes a random sample of 15 games she has bowled and finds the sample mean to be 172. She knows that the standard deviation of her score is 8. Construct a 99 % confidence interval for her score.
55. Flip a coin 50 times and record the number of tails. Construct a 99 % confidence interval for the proportion of tails in the tossing of a coin.
56. A random sample of 450 people who took Dollar Dave's CPA review course reveals that 310 of them passed the CPA exam on the first try. Construct a 90 % confidence interval for the proportion of people who pass the CPA exam on the first try after taking Dollar Dave's course.
57. A random sample of 225 people who went to the Matchmaker Dating Service finds that 100 of those people found their spouse through the service. Construct

a 95 % confidence interval for the proportion of people who find a spouse through this dating service.

58. A random sample of 200 observations from a population yielded the following summary statistics:

$$\sum x = 1,202 \quad \sum x^2 = 121,020$$

Construct a 90 % confidence interval for the population mean μ..

59. A random sample of 80 observations from a population yielded the following summary statistics:

$$\sum x = 475 \quad \sum (x - \bar{x})^2 = 772$$

Construct a 95 % confidence interval for the population mean μ.

60. A random sample of 100 bullets in a case of 1,000 includes 5 that are defective. Construct a 99 % confidence interval for the proportion of defective bullets in a case.
61. Suppose a golfer on the University of Houston golf team plays 70 rounds of golf and breaks par 32 times. Construct a 90 % confidence interval for the proportion of rounds in which this golfer will break par.
62. You roll a die 100 times and get the following results.

Number on die	Number of rolls
1	13
2	16
3	15
4	14
5	22
6	20

Construct a 90 % confidence interval for the proportion of rolls that will be 1 s.

63. Use the information given in question 62 to construct a 90 % confidence interval for the proportion of rolls that will come up 6.
64. A surge in health insurance premiums imposes an additional burden on a business. A random sample of 10 employees indicates that the average cost increase per employee is about $2,345 with a standard deviation of $245. Assuming a normal distribution for the per-employee increase, construct a 90 % confidence interval for the average increase.
65. The owner of a local bakery feels that too many bagels are thrown out every night, so he decides to estimate the demand for bagels. After a month's observation, he collected 30 days' sales and ascertained that the average sales were 120 and the standard deviation of daily sales was 10. Assume that the daily bagel sales follow a normal distribution. Construct a 90 % confidence interval for the demand for bagels.

66. Suppose the owner in question 65 observed the sales for 60 days and found the average sales to be 115 with a standard deviation of 12. Obtain the 90 % confidence interval for the demand for bagels.
67. The manager in the local shoe factory wants to estimate the productivity of the midnight shift. He draws a random sample of 10 nights and records the productivity as follows:

124	124	145	132	123	124	122	141	133	122

(a) Estimate the average productivity.
(b) Assuming that the data follow a normal distribution, derive a 95 % confidence interval.

68. A local dairy farm has just installed a new machine that pumps milk into 16-oz bottles. The manager of the farm wants to make sure that the amount of milk put in the bottles is 16 oz, so he randomly selects 12 bottles of milk each hour and weighs the milk. The results obtained in the last hour were

16.01	16.03	15.89	15.99	16.02	16.03
16.04	16.01	15.99	16.03	16.04	16.05

(a) Obtain the average weight of the milk.
(b) Obtain a 95 % confidence interval for the average amount of milk in the bottles.

69. The personnel office found that in the last 5 years, the average cost of recruiting management trainees has been $500. The cost varies but follows a normal distribution. The standard deviation is estimated to be 25. Assume that the cost of recruitment will remain the same next year and that the company will hire 50 new employees. How much money should the company allocate for recruitment? Construct a 90 % confidence interval to estimate the recruitment expenses.
70. A survey wherein 90 employees were randomly drawn shows that the average number of sick days taken by employees each year is 5.4 days. The number of sick days follows a normal distribution with a standard deviation of 1.5. Obtain a 90 % confidence interval for the average number of sick days.
71. A recent poll shows that 53 % of the voters interviewed strongly support the incumbent and are willing to vote for her in the coming election. The poll was taken by asking 1,000 voters. Estimate the proportion of support for the incumbent with a 95 % confidence interval.
72. A consumer rights organization tests a new car to estimate the car's average gasoline mileage. Because its budget is limited, the organization can test only 25 cars. The standard deviation of the cars tested is 2. What is the range of the 90 % confidence interval?
73. A poll is conducted to predict whether new municipal bonds should be issued. Assume that 230 out of 500 interviewees voted for issuance of the new bonds. How precise is this prediction? Construct a 95 % confidence interval for the proportion of yes votes.

74. In question 73, assume that 45 % of the entire population of voters support issuance of the new bonds. Under this condition, if the pollsters want to stay within 2 % error (plus and minus 1 %), how many voters should they interview?
75. A multinational company wants to find out how society perceives it. The company sends a questionnaire to 2,000 people and learns that 893 have favorable opinions and others either have an unfavorable opinion or no opinion.

    (a) What percent of the people surveyed have favorable opinions of the company? Construct a 90 % confidence interval.
    (b) What is the percentage of people who have favorable opinions of the company? Construct a 95 % confidence interval.

76. When we construct a 90 % confidence interval for, say, a mean, we build a range that has an upper bound and a lower bound, and we write the confidence interval as
    $P(\text{lower bound} < \text{mean} < \text{upper bound}) = 90\ \%$
    Comment on the following statement: would you say the probability that the mean occurs between the upper and lower bounds is 90 %?
77. A new machine was designed to cut a metal part at a length of .24 in.. Although the machine is well designed, for some uncontrolled reasons, the machine cuts the metal with a standard deviation of .01 in.. For quality control purposes, the company wants to draw a sample from each hour's production and measure the average length of the sample metal parts. If the company wants to control the 99 % confidence interval in a range of .01, how many parts should the company sample every hour?
78. The trains scheduled to arrive at the New Brunswick train station at 7:35 A.M. every weekday do not always arrive at 7:35. A commuter carefully recorded the arrival time for the last 200 working days and found that late arrivals follow a normal distribution with a mean delay of 0 min and a standard deviation of 1 min.

    (a) Estimate the average arrival time for the train.
    (b) Estimate the average arrival time using a 90 % confidence interval.
    (c) If you plan to arrive at the train station at 7:34 regularly for the next 200 working days, how many trains should you expect to miss?

79. A marketing consulting company wants to estimate the percentage of students holding credit cards by sending questionnaires to students. The sponsor of this research wants to establish a 95 % confidence interval and a $\pm$ 1 % error margin. To achieve this precision, how many questionnaires should the company send out if every student responds?
80. In a survey of 2,000 voters, 36 % were found to support increasing taxes to build a new school system. Obtain the 95 % confidence interval for the proportion supporting the tax increase.
81. The manager in the local supermarket wanted to know whether it is worth the trouble to keep the store open 24 h a day. He randomly sampled and recorded 20 nights' sales and got

245	145	123	178	125	175	182	130
214	192	120	187	163	148	198	192
129	134	139	271				

Use M1NITAB to answer the following questions:

(a) Estimate the average sales per night.
(b) Construct the 90 % confidence level for the average sales.

82. A large mail-order company wants to find the effect of sending catalogs to potential customers. Of the 600 potential customers who have just received the new catalogs, 123 responded with an order within a month. Estimate the proportion of responses and establish a 90 % confidence interval.
83. The personnel department wants to estimate the cost of hiring a new secretary. The following data are collected on 8 new secretaries:

$2,100	$2,135	$2,545	$2,433
$2,344	$2,564	$2,457	$2,556

Estimate the average cost of hiring. Construct a 90 % confidence interval.
84. The dean of student activities wants to estimate the average spending on beer per week by a student. From a previous study, the standard deviation of spending was estimated to be $39. If the dean wants to control the 90 % confidence interval within $\pm$ $5, how many students should he survey?
85. The dean of student activities wants to know students' reaction to the new student center. Of the 500 students queried, 350 report that they like the new building. Estimate the proportion of the students who like the building. Construct a 90 % confidence interval.
86. In question 85, if the dean wants to narrow the 90 % confidence interval to $\pm$ 1 %, how many students should he ask?
87. A soft drink producer installs a new assembly line to fill 12-oz soda cans. After a week of operation, the plant manager randomly samples 120 cans of soda and weighs the soda. He finds that the soda cans contained an average of 12.05 oz of soda. The standard deviation of the weight is .02 oz. Construct a 95 % confidence interval for the average amount of soda pumped into the cans.
88. In question 87, what is the 95 % confidence interval for the variance of the soda pumped into the cans?

Use the following information to answer questions 89 to 91. In an airline company, a committee was formed to study the seriousness of late arrivals of freight. The following report was compiled about the arrival record:

Total number of freight shipments	625
Total number of late arrivals	159
Average late time	34 min
Standard deviation of late time	25 min

89. Construct a 90 % confidence interval to estimate the average late time.
90. Construct a 90 % confidence interval to estimate the percentage of late arrivals.
91. Construct a 90 % confidence interval to estimate the standard deviation of late time.
92. A potential candidate in the third borough conducted a poll to decide whether he should challenge the incumbent. From a previous poll, he knows that the current incumbent has the support of 45 % of the people. He wants to construct a 90 % confidence interval with a ± 3 % error margin. How many voters should he survey?

Use the following information to answer questions 93–96. An automobile manufacturer wants to study the repair record of its own cars. The performance of 1,000 cars and their maintenance records were monitored after they were sold to consumers. In a span of 3 years, 3,560 repairs occurred among the 1,000 cars monitored. The standard deviation of the number of repairs for 1 car is 2.5. A total of $303,000 was spent to repair the cars. The standard deviation of repair costs for 1 car is $60. There are 205 cars that did not have any repairs in the 3 years.

93. Compute the average cost of 1 repair. Construct an 80 % confidence interval.
94. Compute the average number of repairs for each car. Construct an 80 % confidence interval.
95. Construct a 90 % confidence interval for the standard deviation of costs.
96. Construct a 90 % confidence interval for the proportion of trouble-free cars. What is the error margin?
97. Define the following:

    (a) Convenience lot
    (b) Single-sampling plan
    (c) Double-sampling plan
    (d) Upper control limit
    (e) Lower control limit
    (f) Acceptance sampling

98. Discuss the similarities and differences among $\bar{X}$ -charts, $\bar{R}$ -charts, S-charts, and P-charts.
99. Thirty samples of 100 items each were inspected, and 68 were found to be defective. Compute control limits for a P-chart.
100. The following table gives the fraction defective for an automotive piston for 20 samples. Three hundred units are inspected each day. Construct a P-chart and interpret the results.

Sample	Fraction Defective	Sample	Fraction Defective
1	0.11	11	0.16
2	0.16	12	0.25
3	0.12	13	0.15

(continued)

(continued)

Sample	Fraction Defective	Sample	Fraction Defective
4	0.10	14	0.12
5	0.09	15	0.11
6	0.12	16	0.11
7	0.12	17	0.14
8	0.15	18	0.18
9	0.09	19	0.10
10	0.13	20	0.13

101. One hundred insurance forms are inspected daily over 25 working days, and the numbers of forms with errors are recorded below. Construct a *P*-chart.

Day	Number Defective	Day	Number Defective
1	4	14	4
2	3	15	1
3	3	16	3
4	2	17	4
5	0	18	0
6	3	19	1
7	0	20	1
8	1	21	0
9	6	22	2
10	3	23	6
11	2	24	2
12	0	25	1
13	0		

102. The monthly incomes (in $1,000) from a random sample of eight workers in a factory are 4.2, 5.1, 7.8, 6.2, 8.2, 5.5, 6.7, and 9.1. Assume the population has a normal distribution. Give your answer for (a) and (b).

(a) Compute the standard error of the sample mean (in dollars).
(b) Compute a 95 % confidence interval for the mean of the population.

103. A machine produces parts used in cars. A sample of 25 parts was taken. The average length in the sample was 15.95 in. with a sample variance of 0.4 in..

(a) Construct a 95 % confidence interval for the population variance.
(b) Construct a 99 % confidence interval for the population variance.

104. A production process is considered in control if no more than 3 % of the items produced are defective. Samples of size 500 are used for the inspection process.

(a) Determine the standard error of the sample proportion.
(b) Determine the upper and the lower control limits for the P-chart.

105. A filling machine is set up to fill bottles with 35 oz of coke each. The standard deviation s is known to be 1.2 oz. The quality control department periodically selects samples of 20 bottles and measures their contents. Assume the distribution of filling volumes is normal.

    (a) Determine the upper and lower control limits and explain what they indicate.
    (b) The means of six samples were 37.8, 29.2, 41.9, 25.9, 32.1, and 43.8 oz. Construct an X bar chart and indicate whether or not the process is in control.

Appendix 1: Control Chart Approach for Cash Management

The Miller–Orr model for cash management starts with the assumption that there are only two forms of assets: cash and marketable securities.[16] It also allows for cash balance movement in both positive and negative directions and for the optimal cash balance to be a range of values rather than a single point estimate. In other words, the Miller–Orr model uses the control chart approach we discussed in Sect. 10.9 to do cash management. This model is especially useful for firms that are unable to predict day-to-day cash inflows and outflows.

Figure 10.20 shows the functioning of the Miller–Orr model. Note that the cash balance is allowed to meander undisturbed as long as it remains within the predetermined boundary range shown by the upper limit H and the lower limit L. At point B, however, the cash balance reaches the maximum allowable level. At this point, the firm could purchase marketable securities in an amount equal to the dashed line, which would lower the cash balance to the "return point" from which it would again be allowed to fluctuate freely. At point M, the firm's cash balance reaches the minimum allowable level. At this point, the firm could sell marketable securities to investors or borrow to bring the cash level back up to the return point.

In how much of a range (H–L) should the cash balance be allowed to fluctuate? According to the Miller–Orr model, the higher the day-to-day variability in cash flows and/or the higher the fixed-transactions cost associated with buying and selling securities, the farther apart the control limits should be set. On the other hand, if the opportunity cost of holding cash (the interest foregone by *not* purchasing marketable securities) is high, the limits should be set closer together. Management's objective is to minimize total costs associated with holding cash. Minimization procedures establish that the spread between the upper and lower cash limit (S), the return point (R), the upper limit (H), and the average cash balance (ACB) are

[16] This section on Miller and Orr's model for cash management is taken from Cheng F. Lee and Joseph E. Finnerty (1990), *Corporate Finance*: *Theory*, *Method*, *and Applications* (New York: Harcourt) pp. 595–598.

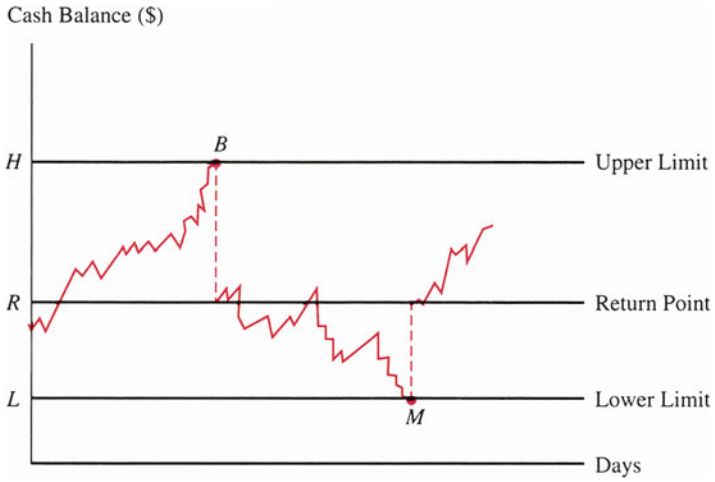


Fig. 10.20 The workings of the Miller–Orr Model (*Source*: Cheng F. Lee and Joseph E. Finnerty, *Corporate Finance*: *Theory*, *Method and Applications*, Fig. 20.3, p. 595. Copyright © 1990 by Harcourt Brace Jovanovich, Inc. Reprinted by permission of the publisher, Harcourt Brace Jovanovich, Inc.)

$$S = 3\left[\frac{3F\sigma^2}{4k}\right]^{1/3} \tag{10.22}$$

$$R = \frac{S}{3} + L \tag{10.23}$$

$$H = S + L \tag{10.24}$$

$$\text{ACB} = \frac{4R - L}{3} \tag{10.25}$$

where L = lower limit, F = fixed-transactions cost, k = opportunity cost on a daily basis, and α^2 = variance of net daily cash flow.

The firm always returns to a point one-third of the spread between the lower and upper limits. With the return point set here, the firm is likely to bump against its lower limit more frequently than against its upper limit. Although this lower point does not minimize the number of transactions and their resulting cost (as the middle point would), it is an optimal point in that it minimizes the sum of transactions cost and foregone-interest cost, the latter of which the firm incurs whenever it holds excessive cash.

To use the Miller–Orr model, the financial manager takes three steps:

1. Set the lower limit.
2. Estimate the variance of cash.
3. Determine the relevant transactions cost and lost-interest cost.

Setting the lower limit is essentially a subjective task, but common sense and experience help. The lower limit is likely to be some minimum safety margin above 0. An important consideration in setting this limit is any bank requirements that must be satisfied.

To estimate the variance of cash flows, the manager can record the net cash inflows and outflows for each of the preceding 100 days and then compute the variance of those 100 observations. This approach requires regular updating, particularly if net cash flows have been unstable over time. One additional aspect to consider in this calculation is the impact of seasonal effects (see Chap. 18), which may also require adjusting the variance estimate.

To determine the relevant transactions cost, the financial manager need only observe what the firm currently pays to buy or sell a security. Interest foregone can be derived from current available market returns on short-term, high-grade securities. The financial manager may want to use a forecasted interest rate for the planning period if a significant change from current interest-rate levels is expected.

We now demonstrate the actual calculations for the Miller–Orr model. First, assume the following:

Minimum cash balance = \$20,000
Variance of daily cash flows = \$9,000,000 (hence the standard deviation σ = \$3,000 per day)
Interest rate = 0.0329 % per day
Transactions cost (average) of buying or selling one security = \$20

Utilizing these data, we first compute the spread between the lower and upper limits in accordance with Eq. 10.22:

$$\begin{aligned}\text{Spread} &= 3\left(\frac{3 \times 20 \times 9,000,000}{4 \times .000329}\right)^{1/3} \\ &= \$22,293\end{aligned}$$

Next, we compute the upper limit and return point in accordance with Eqs. 10.24 and 10.23:

$$\begin{aligned}\text{Upper limit} &= \text{lower limit} + \text{spread} \\ &= \$20,000 + \$22,293 \\ &= \$42,293 \\ \text{Return point} &= 20,000 + \left(\frac{22,293}{3}\right) \\ &= \$27,431\end{aligned}$$

Using Eq. 10.25, we find the average cash balance:

$$\begin{aligned}\text{Average cash balance} &= \frac{4(\$27,431) - \$20,000}{3} \\ &= \$29,908.\end{aligned}$$

Then, on the basis of our assumed input values and model calculations, we can establish the following rule:

If the cash balance rises to \$42,293, invest \$42,293–\$27,431 = \$14,862 in marketable securities; if the cash balance falls to \$20,000, sell \$27,431–\$20,000 = \$7,431 of marketable securities. Both will restore the cash balance to the return point.

Appendix 2: Using MINITAB to Generate Control Charts

This appendix shows how MINITAB may be used to generate an $\bar{X}$ -chart, an $\bar{R}$ – chart, and an S-chart based on the following data:

```
MTB > SET INTO C1
DATA> 10.65 10.70 10.65 10.65 10.85
DATA> 10.75 10.85 10.75 10.85 10.65
DATA> 10.75 10.80 10.80 10.70 10.75
DATA> 10.60 10.70 10.70 10.75 10.65
DATA> 10.70 10.75 10.65 10.85 10.80
DATA> 10.60 10.75 10.75 10.85 10.70
DATA> 10.60 10.80 10.70 10.75 10.75
DATA> 10.75 10.80 10.65 10.75 10.70
DATA> 10.65 10.80 10.85 10.85 10.75
DATA> 10.60 10.70 10.60 10.80 10.65
DATA> 10.80 10.75 10.90 10.50 10.85
DATA> 10.85 10.75 10.85 10.65 10.75
DATA> 10.70 10.70 10.75 10.75 10.70
DATA> 10.65 10.70 10.85 10.75 10.60
DATA> 10.75 10.80 10.75 10.80 10.65
DATA> 10.90 10.80 10.80 10.75 10.85
DATA> 10.75 10.70 10.85 10.70 10.80
DATA> 10.75 10.70 10.60 10.70 10.60
DATA> 10.65 10.65 10.85 10.65 10.70
DATA> 10.60 10.60 10.65 10.55 10.65
DATA> 10.50 10.55 10.65 10.80 10.80
DATA> 10.80 10.65 10.75 10.65 10.65
DATA> 10.65 10.60 10.65 10.60 10.70
DATA> 10.65 10.70 10.70 10.60 10.65
DATA> END
MTB > PAPER
```

Step 1: We input the data into MINITAB, storing it in Column 1 (C1) as presented in the data above.

Step 2: We can use different commands to ask MINITAB to generate an $\bar{X}$ -chart, an $\bar{R}$ *-chart*, or an S-chart.

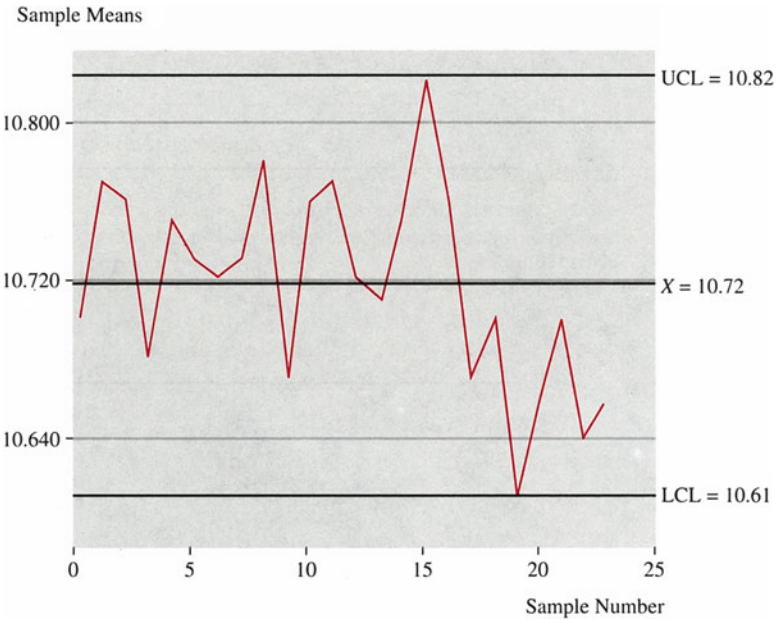


Fig. 10.21 $\bar{X}$ -chart for C1

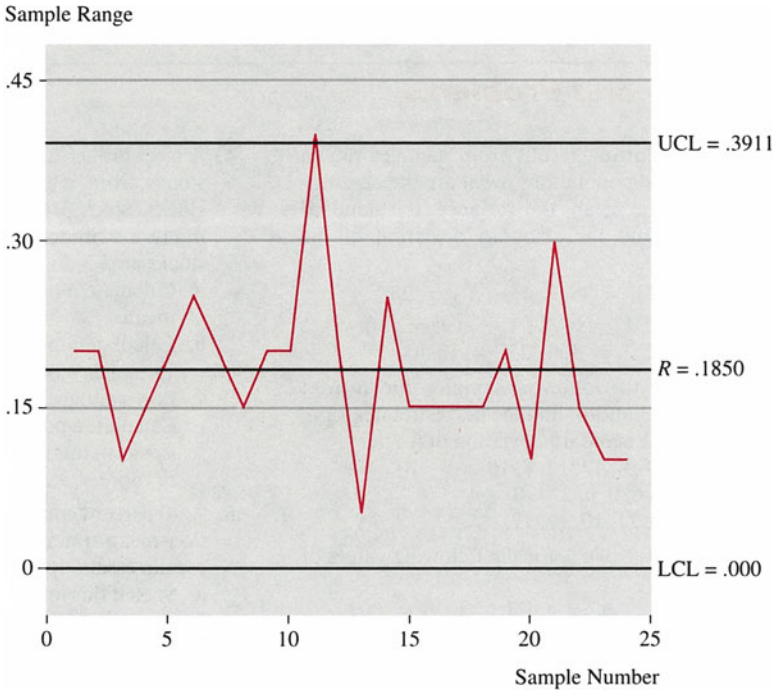


Fig. 10.22 $\bar{R}$ -chart for C1

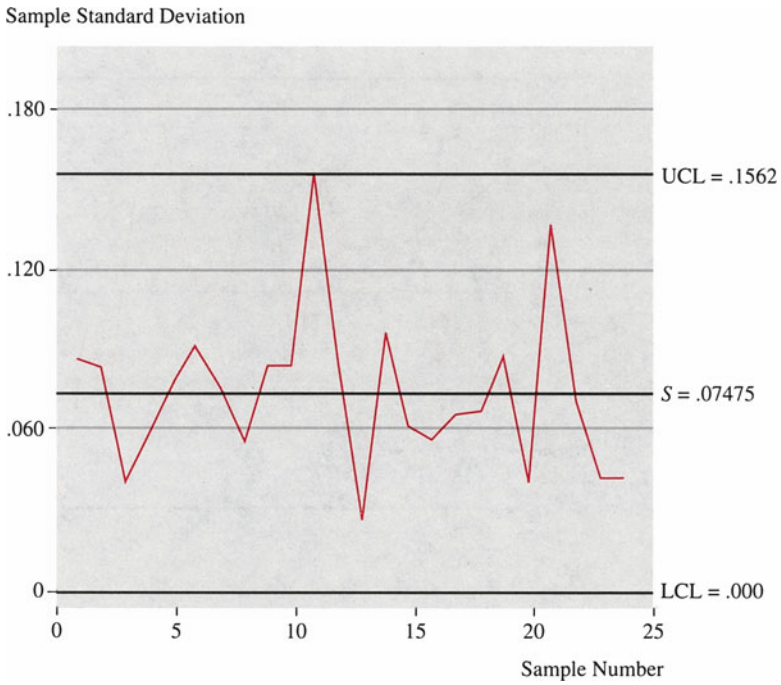


Fig. 10.23 *S*-chart for C1

The command to generate the $\bar{X}$-chart is "XBARCHART C1 5," the command to generate the $\bar{R}$-chart is "RCHART C1 5," and the command to generate the *S*-chart is "SCHART C1 5."

In these three commands, "C1" indicates where the data is located and "5" indicates that there are five observations in each sample. The output for the $\bar{X}$-chart, the $\bar{R}$-chart, and the S-chart is presented in Figs. 10.21, 10.22, and 10.23.

Chapter 11
Hypothesis Testing

Chapter Outline

Key Terms

Hypotheses	Composite hypothesis
Hypothesis testing	Probability value (p-value)
Null hypothesis	Observed level of significance
Alternative hypothesis	The power of a test
Mutually exclusive	Chi-square test
Exhaustive	Power function
Type I error	Power curve
Type II error	Operating-characteristic curve (OC curve)
One-tailed test	Acceptance sampling
Two-tailed test	Operating characteristic
Acceptance region	Lot tolerance percentage defective (LTPD)
Rejection region	Consumer's risk
Lower-tailed test	Acceptable quality level (AQL)
Upper-tailed test	Producer's risk
Critical value	
Simple hypothesis	
Parameter	

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_11, © Springer Science+Business Media New York 2013

11.1 Introduction

Business managers must always be ready to make decisions and take action on the basis of available information. During the process of decision making, managers form hypotheses that they can scientifically test by using that available information. They then make decisions in the light of the outcome. In this chapter, we use the concepts of point estimate and interval estimate discussed in Chaps. 8, 9, and 10 to test hypotheses made about population parameters on the basis of sample data.

Hypotheses are assumptions about a population parameter. *Hypothesis testing* involves judging the correctness of the hypotheses. In fact, we often rely heavily on sample data in decision making. For example, the results of public opinion polls may actually dictate whether a presidential candidate decides to keep running or to drop out of the primary race. Similarly, a firm may use a market sampling survey to gauge consumer interest in a given product and thus determine whether to allocate funds for research and development of that product. And a plant manager may use a sample of canned food products produced by a food canning machine to determine whether the quality of this year's products is the same as that of the previous year's offering.

In this chapter, we first discuss the basic concepts of hypothesis testing and the errors it is subject to. Second, methods of constructing a hypothesis test and testing procedures are explored. Third, we examine in detail one-tailed tests and two-tailed tests for large samples. Small-sample hypothesis tests for means and chi-square tests of a normal distribution variance are discussed next. Then we investigate hypothesis testing for a population proportion and compare the variances of two normal populations. Finally, we present some business applications of hypothesis testing. The power of a test, the power function, and the operating-characteristic curve are discussed in Appendix 1.

11.2 Concepts and Errors of Hypothesis Testing

11.2.1 Concepts

The information obtained from the sample can be used to make inferential statements about the characteristics of the population from which the sample is drawn. One way to do this is to estimate unknown population parameters by calculating point estimates and confidence-interval estimates.

Alternatively, we can use sample information to assess the validity of a hypothesis about the population. For example, the production manager in charge of a cereal box filling process hypothesizes that the average weight of a box of cereal is 30 ounces.

In statistics, hypotheses always come in pairs: the null hypothesis and the alternative hypothesis. The statistical hypothesis that is being tested is called the *null hypothesis*. Our cereal production manager can use a sample of 35 boxes and calculate their average weight and variance to ascertain the validity of the following null and alternative hypotheses:

1. The average weight of cereal per box is 30 ounces (the null hypothesis).
2. The average weight of cereal per box is not 30 ounces (the alternative hypothesis). This implies that it is less than 30 ounces or it is more than 30 ounces.

Rejection of the null hypothesis that is tested implies acceptance of the other hypothesis. This other hypothesis is called the *alternative hypothesis*. These two hypotheses represent mutually exclusive and exhaustive theories about the value of a population parameter such as population mean μ, population variance σ^2, or population proportion P. When hypotheses are *mutually exclusive*, it is impossible for both to be true. When they are *exhaustive*, they cover all the possibilities, that is, either the null hypothesis or the alternative hypothesis must be true.

The null hypothesis is traditionally denoted as H_0, and the alternative hypothesis as H_1. Each of these symbols is always followed by a colon and then by the statement about a population parameter.

The first problem we encounter in hypothesis testing is how to construct the test. To construct a hypothesis test, we first need to specify the null and alternative hypotheses. Because our goal in hypothesis testing is to find out whether we can reject the null hypothesis, we set up the null hypothesis so that it is consistent with the status quo. In addition, H_0 has to be a specific value so that the sampling distribution under H_0 can be determined for the test. By constructing our hypothesis test in this way, we ensure that the status quo is maintained unless we have sufficient information to prove otherwise (i.e., unless we are able to reject H_0). For example, to minimize the risk of sending an innocent person to jail, our legal system is set up so that the accused is "innocent until proven guilty." We can restate this principle as the following null and alternative hypothesis:

$$\begin{aligned} H_0 &: \text{Not guilty} \\ H_1 &: \text{Guilty} \end{aligned} \tag{11.1}$$

The hypothesis test is set up in this way so that the status quo (innocence) is upheld unless the test results show "beyond a reasonable doubt" that the null hypothesis should be rejected.

Another example of hypothesis testing is testing whether the average weight of a package of cookies is equal to the required weight. In this case, the hypotheses are

$$\begin{aligned} H_0 &: \mu = w^* \\ H_1 &: \mu \neq w^* \end{aligned}$$

where w^* is the required weight for each pack of cookies. The manufacturer does not want more or less than the required amount of cookies in each package.

Table 11.1 Actions and the states of nature of the null hypothesis

	State of nature	
Action	*H_0 is true*	*H_0 is false*
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Therefore, it conducts a test to determine whether the statistics from the sample show any severe deviation from the required weight. If H_0 is rejected, then the manufacturer must impose tighter control on the packing process.

11.2.2 Type I and Type II Errors

It seems like a very simple idea that if we can't reject the null hypothesis, we accept it. But we must think twice before accepting H_0. When H_0 cannot be rejected, there are two possibilities: (1) H_0 is indeed true and (2) H_0 is wrong anyway. Maybe the sample size was not large enough or, for some other reason, test results did not enable us to reject H_0. In any case, we cannot conclude from the fact that H_0 can't be rejected that H_0 is necessarily true. We can only say that, on the basis of the sample under study, we can't reject the null hypothesis. For example, suppose we are interested in testing the null hypothesis that there is no life on Mars. It is clear from this example that we will never be able to show that this statement is true. Why? Because even if astronauts are unable to find life on Mars, it doesn't mean that there are no living things on Mars—only that the astronauts were unable to find any living things. However, it *will* be possible to reject the null hypothesis if these astronauts *do* find life on Mars. In other words, we can never prove that the null hypothesis is true but only that we are able or unable to reject it.

Table 11.1 illustrates the relationship between the actions we take concerning a null hypothesis and the truth or falsity of that hypothesis (which is called the state of nature). This table shows that the errors made in testing hypotheses are of two types. We make a *Type I error* when H_0 is true, but we reject it. We make a *Type II error* when H_0 is false, but we accept it.

11.3 Hypothesis Test Construction and Testing Procedure

11.3.1 Two Types of Hypothesis Tests

There are two types of hypothesis testing that we will be interested in: (1) testing whether or not the population mean is equal to a specific value (including zero) and (2) testing whether the population mean is greater than (or less than) a specific value. The first test is a *two-tailed test*; the other is a *one-tailed test*. These two concepts and the related testing procedures will be discussed in detail in Sects. 11.4 and 11.5.

The first step in our hypothesis testing procedure is to divide the sample space into two mutually exclusive areas, the *acceptance region* and the *rejection* (or critical) *region*. We begin by assuming that we have a large sample ($n > 30$) so that we can use the central limit theorem. Later we will examine how our hypothesis testing procedure can be modified to account for small samples ($n < 30$). Where the acceptance and rejection regions lie depends on two things: whether the test is a one- or a two-tailed test and the significance level we assign to our test. The significance level, α, refers to the size of the Type I error that we are willing to accept. In other words, α represents the probability of Type I error:

$$\begin{aligned}\alpha &= P(\text{reject } H_0 | H_0 \text{ is correct}) \\ &= P(\text{Type I error})\end{aligned}$$

Similarly, the probability of Type II error can be defined as

$$\begin{aligned}\beta &= P(\text{fail to reject } H_0 | H_0 \text{ is false}) \\ &= P(\text{Type II error})\end{aligned}$$

How large a significance level we choose depends on the costs associated with making a Type I error. For example, the significance level used by a cookie company interested in the average weight of a package of cookies should differ from the significance level used by a pharmaceutical company interested in the average amount of an active ingredient in one of its medications. Clearly, the cost to the cookie company of having too many or too few cookies in a package is small compared to the cost to the pharmaceutical company of using too much or too little of an active ingredient in one of its products. (Too little may render the product ineffective; too much may cause the product to lead to harmful side effects or even death.) Similarly, there are costs associated with making a Type II error (failing to reject the null hypothesis even though it is false). The cost associated with making a Type II error is also smaller for the cookie company than for the pharmaceutical firm.

Figure 11.1 illustrates the sampling distribution of the sample mean $\bar{X}$, showing the acceptance and rejection regions for a null hypothesis. Here we display only Type I error. We will discuss both Type I and Type II errors and the trade-off between these two types of errors in the next section.

In Fig. 11.1, C_L is the critical value for the lower-tailed test, and C_U is the critical value for the upper-tailed test. C_L and C_U are the critical values for the two-tailed test. The *critical value* is the cutoff point for hypothesis testing; its value depends on a level of probability, such as 5 percent, 1 percent, or some other percentage.

Figure 11.1a presents the case of a *lower-tailed test*. We conduct a one-tailed test in the lower tail of the distribution when we are concerned only with when the population mean μ is smaller than some specified value μ_0. For example, an investor who is trying to evaluate a stockbroker's performance may be concerned only with below-par performance. In this case, a lower-tailed test is in order, and the investor rejects the null hypothesis of average or above-average performance on the part of the stockbroker if the broker's mean return $\bar{X}$ is less than the critical value C_L.

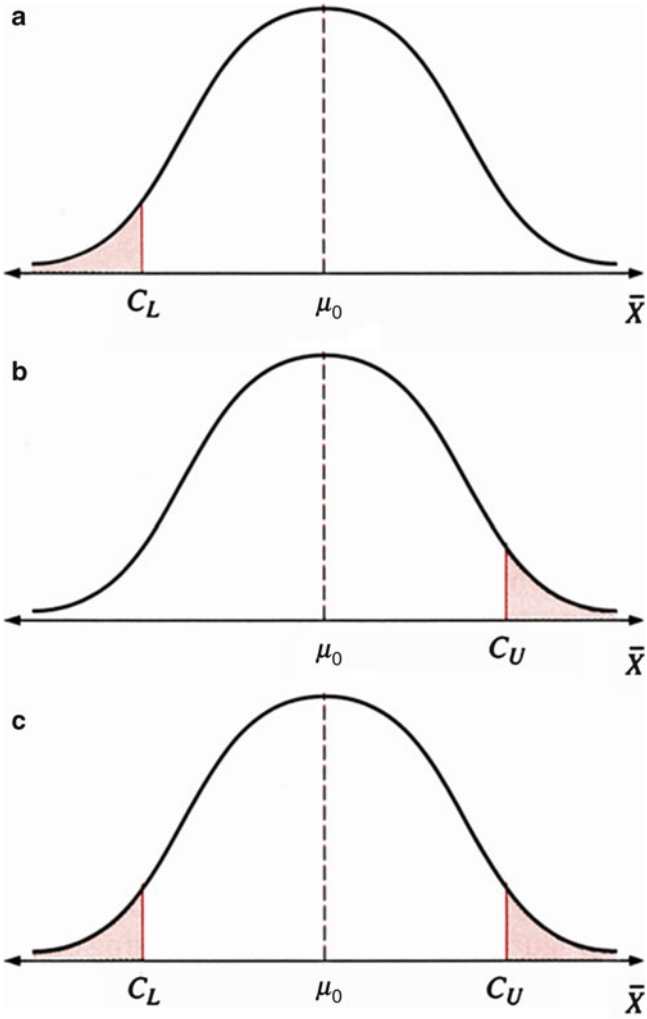


Fig. 11.1 Different types of hypothesis testing: (**a**) lower-tailed test, (**b**) upper-tailed test, and (**c**) two-tailed test

An *upper-tailed test* (Fig. 11.1b) is in order when we are concerned only with when the population mean μ is larger than the specified value μ_0. For example, a pharmaceutical company might be interested in the average amount of an active ingredient in its sleeping pills. Because too much of the active ingredient may lead to harmful side effects, the company may choose to conduct an upper-tailed test. In Fig. 11.1b, we can see that the company rejects the null hypothesis of an average or below-average amount of the active ingredient if the sample mean of the sleeping pills tested, μ, is greater than the critical value C_U.

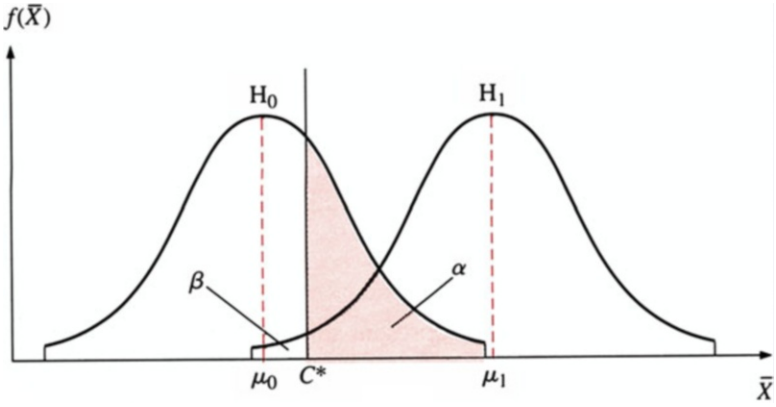


Fig. 11.2 α and β when sample size = n

A two-tailed test is called for when we are interested in the population mean μ being either much larger or much smaller than the specified value μ_0. For example, a cookie manufacturer is interested in the average number of cookies per package. Too few cookies in a package may lead to complaints from consumers; too many will reduce the company's profits. In this case (Fig. 11.1c), the company rejects the null hypothesis of the correct number of cookies in a package if the sample mean $\bar{X}$ falls below the lower critical value C_L or above the upper critical value C_U.

11.3.2 The Trade-off Between Type I and Type II Errors

One way to visualize the trade-off between Type I and Type II errors is to assume that there are only two distributions in which we are interested. One distribution corresponds to H_0, and the other is consistent with H_1. In this case, we are assuming that both the null and alternative hypotheses are simple. A *simple hypothesis* is one wherein we specify only a single value for the population parameter, θ. A *parameter* is a summary measure that is computed to describe a characteristic of an entire population. For example, we might be interested in testing H_0: $\mu = 5$ versus H_1: $\mu = 8$. In this example, both the null and the alternative hypotheses are simple.

On the other hand, we may choose to specify a range of values for the parameter θ. In this case, the hypothesis is called a *composite hypothesis*. For example, we could test a simple null hypothesis, H_0: $\mu = 5$, and a composite alternative hypothesis, H_1; $\mu > 5$. Here, the alternative hypothesis is composite because H_1 is consistent with a range of values for μ. For both simple and composite hypotheses, we need to choose a significance level such as $\alpha = .10, .05$, or $.01$.

In order to present the relationship between Type I and Type II errors in the simplest fashion, let's examine the case of testing a simple null hypothesis, H_0:

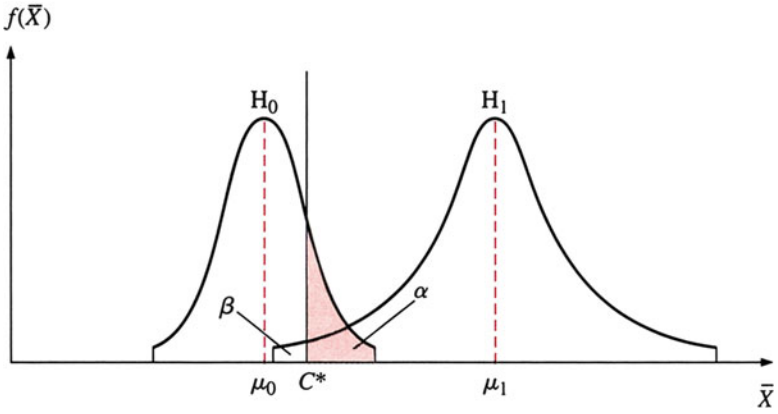


Fig. 11.3 α and β when sample size $= n'(n' > n)$

$\mu = \mu_0$, and a simple alternative hypothesis, H_1: $\mu = \mu_1$. Figs. 11.2 and 11.3 is a graph for this example.

In Fig. 11.2, we can see that there are two distinct locations for the distributions $\bar{X}$. One corresponds to the null hypothesis, and the other to the alternative hypothesis. Because the two distributions overlap, two possible errors can result from our hypothesis test. Type I error occurs when we reject H_0 when it is true. That is, even though the distribution of $\bar{X}$ is consistent with H_0, we reject H_0 because our sample mean is larger than the critical value C^*. Type I error is represented by α, the area under the H_0 distribution curve that lies to the right of the critical value C^*. Type II error occurs when we fail to reject H_0 when H_1, instead of H_0, is correct. In this case, even though the distribution of $\bar{X}$ is consistent with H_1, we accept H_0 because our sample mean is smaller than the critical value C^*. Type II error is represented by β, the area under the H_1 distribution curve that lies to the left of the critical value C^*.

As we can see, the areas α and β are related. If we choose to make α smaller (i.e., reduce the chance of a Type I error), we must settle for a larger β (i.e., increase the chance of a Type II error). This is the trade-off between α and β.

Does this trade-off imply that the only way for us to reduce the chance of making a Type II error is to settle for a larger chance of making a Type I error? The answer to this question is no. By increasing our sample size, it is possible to reduce our chance of making a Type II error without increasing our chance of making a Type I error. The sample standard deviation of both H_0 and H_1 distributions can be defined as

$$s_{\bar{X}} = s_X/\sqrt{n} \tag{11.2}$$

where s_x and n represent sample standard deviation and sample size, respectively. If sample size increases from n to n', then the standard deviations of both H_0 and

H_1 distributions become smaller. Hence, both α and β decrease, as indicated in Fig. 11.3.

The relationship between the critical value C^* and sample size can be written as

$$C^* = \mu_0 + z_0\left(\frac{s_X}{\sqrt{n}}\right) \tag{11.3}$$

$$C^* = \mu_1 - z_1\left(\frac{s_X}{\sqrt{n}}\right) \tag{11.4}$$

where z_0 and z_1 represent the standard deviation units to the right of μ_0 and the standard deviation units to the left of μ_1, respectively.

We can express the required sample size by solving the simultaneous Eqs. 11.3 and 11.4. The solution is

$$n = \left[\frac{(z_0 + z_1)s_X}{(\mu_1 - \mu_0)}\right]^2 \tag{11.5}$$

For example, let $z_0 = 1.60$, $z_1 = 1.80$, $\mu_0 = 550$, $\mu_1 = 580$, and $s_x = 200$. Then the required sample size is

$$n = \left[\frac{(1.60 + 1.80)(200)}{(580 - 550)}\right]^2$$

$$n = 22.67$$

Therefore, a simple random sample of 23 (to the nearest integer) should be required in order to obtain the desired levels of error control. The critical value can be obtained by substituting $n = 23$ into either Eqs. 11.3 or 11.4. Substituting into Eq. 11.3 yields

$$C^* = 550 + (1.60)\left(\frac{200}{\sqrt{23}}\right) = 616.67$$

Applications of Eqs. 11.3, 11.4, and 11.2 in quality control will be discussed in Appendix 1

11.3.3 The P-Value Approach to Hypothesis Testing

Another approach to hypothesis testing is through the use of a *probability value (p-value)*. Under this approach, rather than testing a hypothesis at such preassigned

levels of significance as $\alpha = .05$ or .01, investigators often determine the smallest level of significance at which a null hypothesis can be rejected. The *p*-value is this significance level. In other words, it is the probability of getting a value of the test statistic as extreme as or more extreme than that which is actually obtained, given that the tested null hypothesis is true. Using the *p*-value in hypothesis testing enables us to determine how significant or insignificant our test results are. Did we barely reject the null hypothesis or did we reject it overwhelmingly? The *p*-value is often referred to as the *observed level of significance*. If the *p*-value is smaller than or equal to significance level α, the null hypothesis is rejected; if the *p*-value is greater than α, the null hypothesis is not rejected. The advantage of the *p*-value approach is that it frees us from having to choose a value of α. The disadvantage is that we may obtain an inconclusive test. Applications of the *p*-value will be discussed further in the next two sections.

So far, our discussion of hypothesis testing has focused on determining the level of significance, α, of our test. In addition, we discussed the method of computing a critical value in terms of α and *p*-value and examined the relationship between the *p*-value and α. In all cases, our tests involved controlling the Type I error, α. We have also discussed the trade-off between Type I error and Type II error. It is important to investigate how well the hypothesis test controls Type II errors. The *power of a test,* which is defined as $1–\beta$, can be used to measure how well Type II error has been controlled. This issue and related concepts will be discussed in Appendix 1.

11.4 One-Tailed Tests of Means for Large Samples

As we noted in Sect. 11.3, hypothesis tests can be conducted as one-tailed or two-tailed tests. In this section, we further examine one-tailed tests of means. We begin by examining the case where only one sample is drawn and where that sample is large. Using a large sample offers two important advantages. A large sample makes it possible to apply the central limit theorem. And it enables us, through our choice of significance level (α), to reduce our chance of making a Type II error.

11.4.1 One-Sample Tests of Means

In this section, we examine the one-tailed test of means where only one large random sample is taken. In this case, the null hypothesis is that the population mean is equal to some specified value μ_0. This hypothesis is denoted H_0: $\mu = \mu_0$. Suppose the alternative hypothesis of interest is that the population mean is smaller than this specified value, that is, H_1: $\mu < \mu_0$.

It is natural to base tests of population mean μ on the sample mean $\bar{X}$. In particular, we would like to know whether the observed sample mean is greatly smaller than the specified value of μ_0. To do this, we require the format of a test with

some preassigned significance level α. As described in the previous section, α is used to denote the Type I error.

By using the central limit theorem, we saw in Chap. 8 that when the sample size is large, the sample mean $\bar{X}$ is approximately normally distributed. Therefore, the random variable Z, defined in Eq. 11.6, follows a standard normal distribution:

$$Z = \frac{\overline{X} - \mu_0}{\sigma_X / \sqrt{n}} \tag{11.6}$$

This equation implies that the sampling distribution of the sample mean $\bar{X}$ is normally distributed with mean μ_0 and standard deviation $\sigma_x / \sqrt{n}$ when the null hypothesis is true. For large samples, the sample standard deviation s can be used in place of σ in Eq. 11.6. The null hypothesis is to be rejected if the sample mean $\bar{X}$ is greatly smaller than the hypothesized value μ_0. Thus, we will reject H_0 if we observe a large absolute value of the random variable Z, as indicated in Eq. 11.6.[1]

If the Type I error α is fixed, then we can follow Chap. 10 in using z_α, for which $P(Z < -z_\alpha) = \alpha$. If the null hypothesis is true, then the probability that the random variable as indicated in Eq. 11.6 is smaller than $-z_\alpha$ is α. In terms of sample mean $\bar{X}$, the decision rule is

$$\text{Reject } H_0 \text{ if } \frac{\overline{X} - \mu_0}{\sigma_X / \sqrt{n}} < -z_\alpha$$

This situation is illustrated in Fig. 11.4. In this case, because z_α is in the lower tail, we have a lower-tailed hypothesis test.

Alternatively, by letting

$$\frac{\overline{X}_\alpha - \mu_0}{\sigma_x / \sqrt{n}} = -z_\alpha$$

we can obtain $\bar{X}_\alpha$ as indicated in Eq. 11.7:

$$\overline{X}_\alpha = \mu_0 - Z_\alpha \sigma_{\overline{X}} = \mu_0 - Z_\alpha \sigma_X / \sqrt{n} \tag{11.7}$$

$\bar{X}_\alpha$ can be used as an acceptance limit for performing the null hypothesis test (see Fig. 11.5). From the normal distribution in Table A3 of Appendix A at the end of the book, we find that $P(Z \leq -1.645) = .05$. If $\alpha = .05$, then $\bar{X}_\alpha$ can be estimated as μ_0–(1.645) $\sigma_X / \sqrt{n}$. Similarly, when $\alpha = .01$, we find that $z = -1.96$ and $\bar{X}_\alpha = \mu_0$–(1.96) $\sigma_x / \sqrt{n}$. If $\bar{X}$ is smaller than $\bar{X}_\alpha$, then we reject the null hypothesis.

[1] The observed value of Z is negative if the alternative hypothesis is $\mu < \mu_0$. It is positive if the alternative hypothesis is $\mu > \mu_0$.

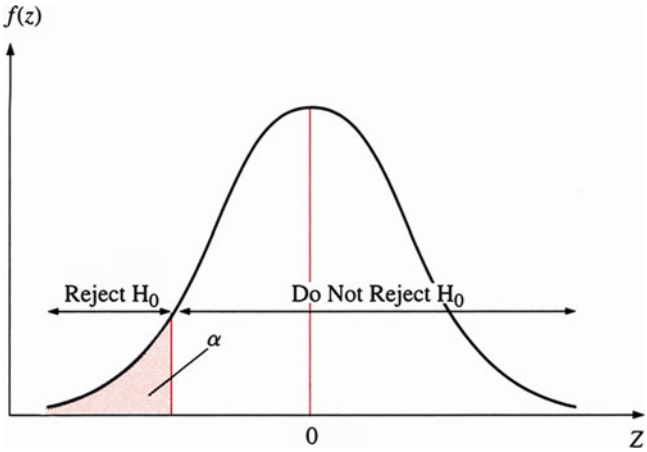


Fig. 11.4 Lower-tailed hypothesis test at the α significance level

Example 11.1 Testing the Average Weight of Cat Food per Bag. Say we want to test whether the average weight of 60-ounce bags of cat food is equal to or smaller than 60 ounces at significance level $\alpha = .05$. The null and alternative hypotheses can be stated as

$$H_0 : \mu = 60$$
$$H_1 : \mu < 60$$

In addition, suppose we know that sample size $n = 100$, sample mean $\bar{X} = 59$, and standard deviation $s_X = 5$. We now will use three different approaches to do the test.

11.4.2 The z_α-Value Approach

Substituting related information into Eq. 11.6, we obtain

$$Z = \frac{59 - 60}{\frac{5}{10}} = -2$$

If $\alpha = .05$, the test statistic is $-z_{.05} = -1.645$, as indicted in Fig. 11.6. A glance at Fig. 11.6 reveals that -2 is in the rejection region, so we reject the null hypothesis.

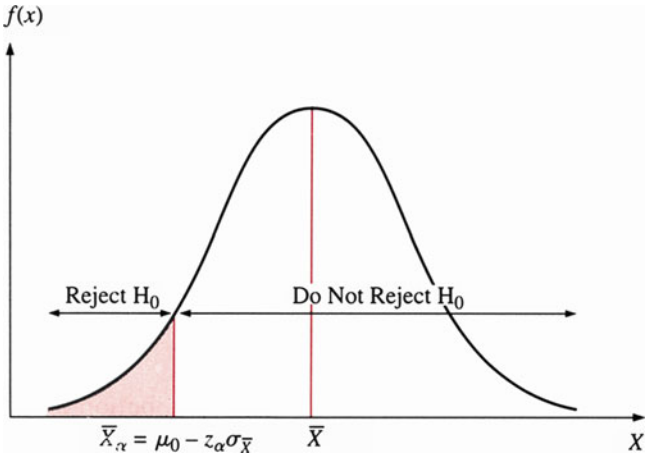


Fig. 11.5 Lower-tailed hypothesis test at the $\bar{X}_\alpha$ significance level

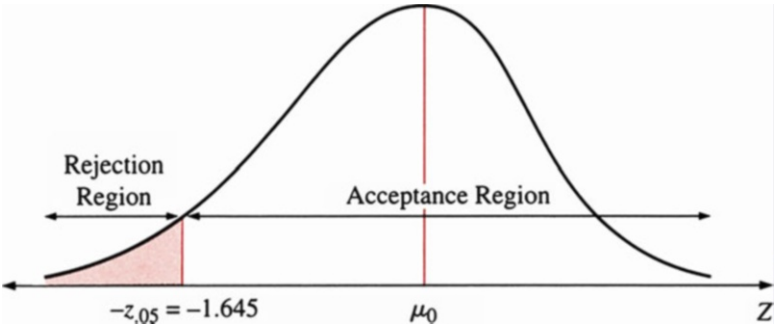


Fig. 11.6 The location of the critical value in a lower-tailed test when the test statistic is $-z_{0.5} = -1.645$

11.4.3 The $\bar{x}_\alpha$ -Value Approach

Substituting this information into Eq. 11.7 in terms of $\alpha = .05$, we obtain $\bar{X}_{.05} = 60-(1.645 \times 5)/10 = 60-8.225/10 = 59.1775$.

Figure 11.7 reveals that the observed sample mean of 59 ounces is in the rejection region, so the null hypothesis, H_0, is rejected.

11.4.4 The p*-Value Approach*

Because this is only a one-tailed test, the *p*-value approach represents the probability in only one tail of the distribution. From the *z*-value approach, we know that

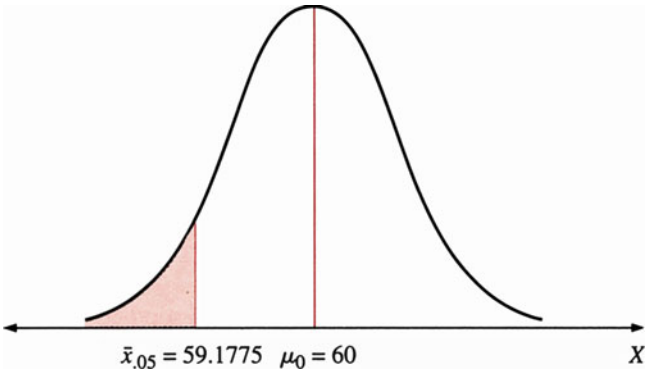


Fig. 11.7 The *location* of the critical value in a lower-tailed test when the test statistic is $\bar{x}_{.05} = 59.1775$

$-z_{.05} = -1.645$. Using the p-value approach, we find that the probability of obtaining a z-value smaller than -2.0 is .500–.4772 = .0228 (see Fig. 11.8). This is less than $\alpha = .05$, so we reject the null hypothesis.

Thus, we can choose among three approaches to doing one-tailed null hypothesis tests. Note that the z_α-value approach is equivalent to the $\bar{x}_\alpha$ -value approach.

11.4.5 Two-Sample Tests of Means

Another important issue is how to test the difference between two population means, μ_1 and μ_2, of two normally distributed populations with variances σ_1^2 and σ_2^2. Because we will use large samples, the assumption of normality is not necessary.

We select two independent random samples from two different populations for n_1 and n_2 observations with observed sample means $\bar{X}_1$ and $\bar{X}_2$. Are we willing to attribute the difference between $\bar{X}_1$ and $\bar{X}_2$ to chance sampling errors, or should we conclude that the populations from which the two samples are drawn have different means? In this case, we have two options: the following one-sided tests with significance level α:

1. Upper-tailed null hypothesis

$$\begin{aligned} \mathrm{H}_0 &: \mu_1 - \mu_2 = D \\ \mathrm{H}_1 &: \mu_1 - \mu_2 > D \end{aligned}$$

where D can be either zero or a positive number.

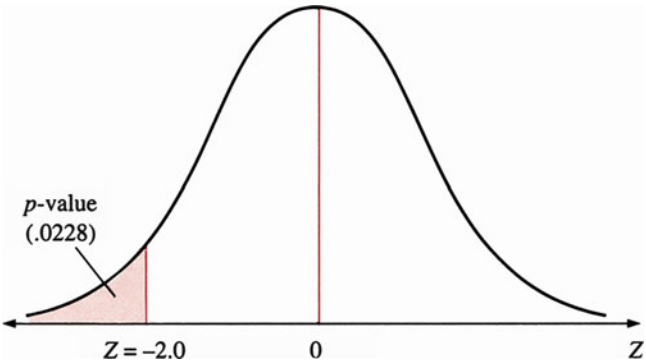


Fig. 11.8 Determining the *p*-value for a one-tailed test

2. Lower-tailed null hypothesis

$$\mathrm{H}_0 : \mu_1 - \mu_2 = D$$
$$\mathrm{H}_1 : \mu_1 - \mu_2 < D$$

where D can be either zero or a positive number.

The z statistic in terms of the central limit theorem that is used to do the aforementioned one-tailed tests can be defined as follows[2]:

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - D}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{11.8}$$

If sample sizes n_1 and n_2 are large, tests of significance level α for the difference between μ_1 and μ_2 are obtained by replacing σ_1^2 and σ_2^2 by s_1^2 and s_2^2. Equation 11.8 can be rewritten as

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{11.9}$$

The following example demonstrates how to test the difference between two population means.

Example 11.2 Comparing Unleaded Gasoline Prices at Texaco and Shell Stations. David Smith conducts a market survey to compare the prices of unleaded

[2] Because $\bar{X}_1$ is independent of $\bar{X}_2$,

$$\mathrm{Var}\ (\bar{X}_1 - \bar{X}_2) = \mathrm{Var}\ (\bar{X}_1) + \mathrm{Var}(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

gasoline at Texaco stations and Shell stations. A random sample of 32 Texaco stations and 38 Shell stations in central New Jersey is used. The cost of 1 gallon of unleaded gasoline is recorded, and the resulting data are summarized here.

Sample A (Texaco)					
1.06	.97	.97	.96	1.02 1.09	
1.08	1.04	1.11	1.12	1.19 1.07	
1.14	1.17	1.22	.97	1.08	
1.05	1.21	.95	.99	1.18	
1.05	1.21	1.03	1.14	1.14	
1.13	1.00	1.16	.96	.98	
	$n_1 = 32$			$\bar{X}_1 = \$1{,}076$	$s_1 = \$.085$
Sample B (Shell)					
1.08	.96	1.06	1.11	1.07	
1.17	1.01	1.05	1.04	1.09	
1.05	1.06	1.14	1.04	.94	
1.01	.99	1.07	1.18	.94	
1.08	1.13	1.16	1.00	.94	
1.13	.91	1.13	.96	.95	
1.00	1.09	1.15	1.13		
.98	1.04	1.03	1.17	.98	
	$n_2 = 38$			$\bar{X}_2 = \$1{,}054$	$s_2 = \$.075$

Is Texaco's average unleaded gasoline price per gallon ($\bar{X}_1$) more than Shell's average price per gallon ($\bar{X}_2$) at $\alpha = .05$? To perform the test, we can follow these steps:

Step 1: Define the hypotheses and evaluate the test statistic.

The question is whether the data support the claim that $\mu_1 > \mu_2$.[3]

H_0: $\mu_1 \leq \mu_2$ (Texaco is less expensive or equally expensive.)

H_1: $\mu_1 > \mu_2$ (Texaco is more expensive.)

From Eq. 11.9, the test statistic can be calculated as

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1.076 - 1.054}{\sqrt{\frac{(.085)^2}{32} + \frac{(.075)^2}{38}}}$$
$$= .022/.0188 = 1.1702.$$

Step 2: Define the rejection region and state a conclusion.

Figure 11.9 indicates that the null hypothesis is to be rejected if $Z > 1.645$ under a significance level of .05. Because $Z = 1.1702$ is smaller than 1.645, we accept H_0, and because $\bar{X}_1$ is not significantly larger than $\bar{X}_2$, we claim that $\mu_1 \leq \mu_2$. We conclude that the Texaco stations charge the same or less for gasoline (on the

[3] This kind of hypothesis is called composite null and alternative hypothesis. The decision rule is identical to that for a simple alternative hypothesis specified as H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 > \mu_2$.

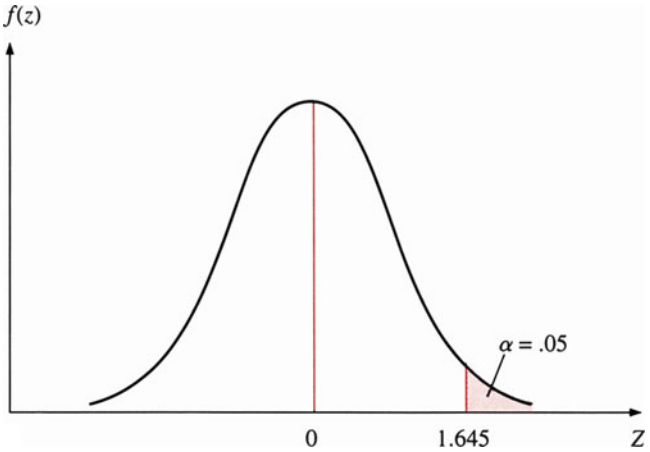


Fig. 11.9 *z* curve showing rejection region for example 11.2

average) than the Shell stations. Alternatively, we input sample data into MINITAB and obtain the mean, standard deviation, *t*-statistic, and *p*-value as follows:

```
MTB >   SET INTO
            C1
DATA>  1.06 1.05  0.97  1.21  0.97  0.95  0.96  0.99  1.02  1.18  1.09
DATA>  1.08 1.05  1.04  1.21  1.11  1.03  1.12  1.14  1.19  1.14  1.07
DATA>  1.14 1.13  1.17  1.00  1.22  1.16  0.97  0.96  1.08  0.98
DATA>     END
MTB >   SET INTO
            C2
DATA>  1.08 1.08  0.96  1.13  1.06  1.16  1.03  1.04  0.96  1.07  0.94
DATA>  1.17 1.13  1.01  0.91  1.05  1.13  1.11  1.18  1.13  1.09  0.94
DATA>  1.05 1.00  1.06  1.09  1.14  1.15  1.04  1.00  1.17  0.94  0.95
DATA>  1.01 0.98  0.99  1.04  1.07
```

```
DATA > END
MTB > TWOSAMPLE C1 C2;
SUBC > ALTERNATIVES = 1.
Two Sample T-Test and Confidence Interval
Two sample T for C1 vs C2
      N        Mean         StDev         SE Mean
C1    32       1.0763       0.0846        0.015
C2    38       1.0537       0.0754        0.012
95% CI for mu C1 - mu C2: ( −0.016, 0.061)
T-Test mu C1 = mu C2 (vs >): T = 1.17 P = 0.12 DP = 62
```

From this computer output, we find that the *t* statistic is equal to 1.17 and the *p*-value equals .12. We conclude, therefore, that the average price per gallon of unleaded gasoline from Texaco stations is the same as that from Shell stations.

11.5 Two-Tailed Tests of Means for Large Samples

11.5.1 One-Sample Tests of Means

A cookie store sells individual cookies and cookies in packages. All the packages are sold for the same price, so the weights of the packages should be equal. If the weight is greater than the specified weight on the packing box, the store suffers a loss. If the weight is less, customers complain. Hence, the store must periodically draw samples and test whether the average weight deviates from the required weight. The hypotheses tested are

$$\begin{aligned} \mathrm{H}_0 &: \mu = D \\ \mathrm{H}_1 &: \mu \neq D \end{aligned}$$

Figure 11.10 illustrates the hypothesis test. Note that the significance level is $\alpha/2$ instead of α.

The decision rule can be either of the following:

1. Reject H_0 if $\bar{x}$ is greater than the upper critical value C_U or less than the lower critical value C_L.
2. Reject H_0 if the p-value is less than α, no matter which tail the sample mean falls in.

Using data from Example 11.1, we set up the two-tailed hypothesis test as

$$\begin{aligned} \mathrm{H}_0 &: \mu = 60 \\ \mathrm{H}_1 &: \mu \neq 60 \end{aligned}$$

and calculate the Z statistic as

$$Z = \frac{59 - 60}{5/10} = -2$$

From the standardized normal distribution table (Table A3 in Appendix A), we know that $z_{\alpha/2} = 1.96$ and $-z_{\alpha/2} = -1.96$. Our $z = -2$ is less than $-z_{\alpha/2}$, so our decision is to reject H_0.

Using the p-value approach, we could determine the probability of obtaining a Z-value smaller than -2.0. From Appendix A, that probability is $.5000 - .4772 = .0228$. Because we are performing a two-tailed test, we also need to find the probability of obtaining a value larger than 2.00. The normal distribution is symmetrical, so this value is also .0228. Thus, the p-value for the two-tailed test is .0456 (see Fig. 11.11). This result may be interpreted to mean that the probability of obtaining a more extreme result than the one observed is .0456. Because this value is smaller than $\alpha = .05$, the null hypothesis is rejected.

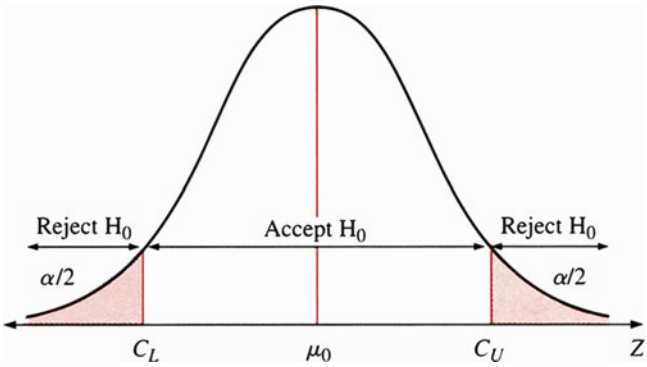


Fig. 11.10 Two-tailed hypothesis testing

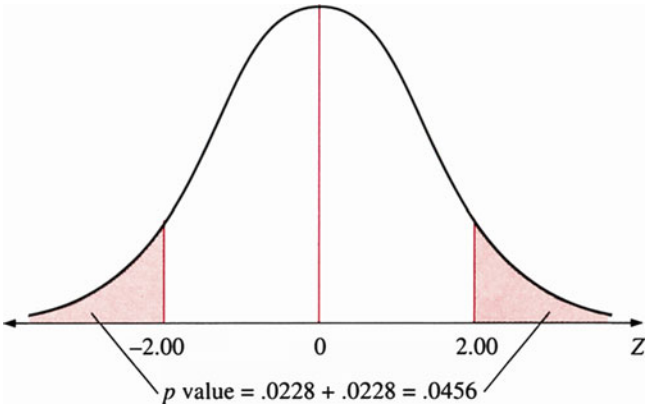


Fig. 11.11 Determining the *p*-value for a two-tailed test

In addition, let's look at this real-world example based on C. S. Patterson's study of a sample of 47 large public electric utilities with revenues of $300 million or more according to Moody's Manual (*Financial Management*, Summer 1984). Patterson's study focused on the financing practices and policies of these regulated utilities. Compilation of the actual debt ratios, or long-term debt divided by total capital, of the companies yielded the following results:

$$\overline{X} = .485 \qquad s_X = .029$$

Before giving their actual debt ratios, the companies cited .459 as the mean debt ratio at which they should operate to maximize shareholder wealth.

From this information, we can test whether the actual mean debt ratio of public utilities differed from the optimum value .459 at $\alpha = .01$. The two-tailed hypothesis test can be defined as

$$\begin{aligned} H_0 &: \mu = .459 \\ H_1 &: \mu \neq .459 \end{aligned}$$

and the z statistic can be calculated as

$$Z = \frac{\overline{X} - \mu_0}{s_X/\sqrt{n}} = \frac{.485 - .459}{.029/\sqrt{47}} = 6.146$$

From Table A3 in Appendix A, we know that $z_{.005} = 2.575$. Since $z > z_{.005}$, we reject the null hypothesis H_0.

11.5.2 Confidence Intervals and Hypothesis Testing

The hypothesis testing discussed in this chapter applies the same concepts as do the confidence intervals we discussed in the last chapter. We used confidence intervals to estimate parameters, whereas we used hypothesis testing to make decisions about specified values of population parameters.

In many situations, we can turn to confidence intervals to test a null hypothesis. This can be illustrated for the test of a hypothesis for a mean. In Example 11.1 (testing whether the average weight of packages of cat food was different from 60 ounces), we employed the formula

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma_X}{\sqrt{n}}} \tag{11.11}$$

We could also solve the cat food problem by obtaining a confidence-interval estimate of μ_0 in terms of sample mean $\bar{x}$. If the hypothesized value of $\bar{x} = 59$ did not fall in the interval, the null hypothesis would be rejected. That is, the value 59 would be considered unusual for the data observed. On the other hand, if it did fall in the interval, the null hypothesis would not be rejected because 59 would not be an unusual value. The confidence-interval estimate in terms of data defined in Example 11.1 was

$$\begin{aligned} &\overline{X} \pm z_\alpha \frac{s_X}{\sqrt{n}} \\ &60 \pm (1.645)\frac{5}{10} = 60 \pm .8225 \end{aligned}$$

so that $59.1775 \leq \mu \leq 60.8225$. This interval does not include the hypothesized value of 59, so we would reject the null hypothesis. This, of course, is the same decision we reached by using the hypothesis testing technique.

11.5.3 Two-Sample Tests of Means

Two-sample tests involve testing the equality of two sample means. Two-tailed tests are similar to one-tailed tests, but the alternative hypothesis H_1 assumes that two population means are "unequal" and the significance level for each tail is now $\alpha/2$. The hypothesis test can be expressed as follows:

$$H_0 : \mu_1 - \mu_2 = D$$
$$H_1 : \mu_1 - \mu_2 \neq D$$

where D can be either zero or a positive number.

In order to test, we can calculate either the p-value or the critical values on both tails. The decision rules are:

1. Reject H_0 if the p-value is less than α.
2. Reject H_0 if $(\bar{X}_1 - \bar{X}_2)$ is either greater than C_U or less than C_L, as shown in Fig. 11.12.

C_L is calculated as follows:

$$C_L = -z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{11.13}$$

If σ_1^2 and σ_2^2 are unknown, sample variances s_1^2 and s_2^2 can be used to approximate C_L, which can be denned as

$$C_L = -z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{11.14}$$

Furthermore,

$$C_U = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

can be approximated by

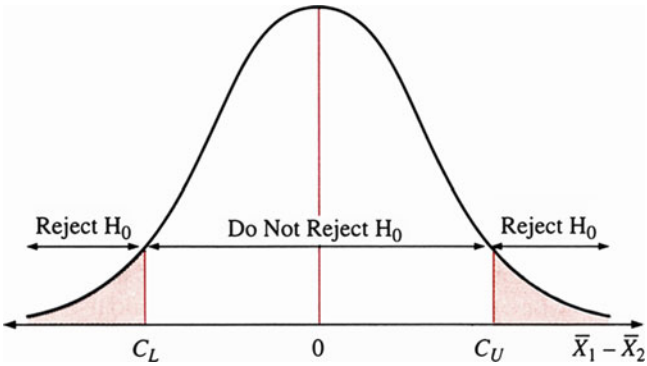


Fig. 11.12 Rejection and acceptance regions for two-samples case

$$z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \tag{11.15}$$

With critical values C_L and C_U, we can perform the null hypothesis test as in the case of a one-sample test.

Using the unleaded gasoline prices in Example 11.2 as an example, we now show how Eqs. 11.14 and 11.15 can be used to do a two-tailed test at $\alpha = .05$. The question is whether data support the claim that $\mu_1 = \mu_2$:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_1 &: \mu_1 - \mu_2 \neq 0 \end{aligned}$$

From Table A3 in Appendix A, we find $z_{.025} = 1.96$. Substituting $z_{.025} = 1.96$, $s_1 = .085$, $s_2 = .075$, $n_1 = 32$, and $n_2 = 38$ into Eqs. 11.14 and 11.15, we obtain

$$\begin{aligned} C_L &= -(1.96)\sqrt{\frac{(.085)^2}{32}+\frac{(.075)^2}{38}} = -(1.96)(.0188) \\ &= -.0368 \\ C_U &= (1.96)(.0188) = .0368 \end{aligned}$$

Since $\bar{X}_1 - \bar{X}_2 = 1.076–1.054 = .022$ is smaller than C_U and larger than C_L, we cannot reject the null hypothesis $\mu_1 = \mu_2$.

11.6 Small-Sample Tests of Means with Unknown Population Standard Deviations

So far, our discussion of hypothesis testing has focused on cases where the sample size is large. This large sample size has enabled us to employ the central limit theorem and to use the normal distribution in our hypothesis tests. If our sample size is small ($n < 30$), however, we must modify our test. As we noted in Chap. 10, when the sample size is small, we should use the t distribution in place of the normal distribution. Note that the use of the t distribution with small samples requires that the original population be distributed normally. Using the Z test for small-sample hypothesis testing leads to inaccurate results. Table 11.2 shows how t_α approaches Z_α as the sample size increases. It gives some idea how "small" a sample should be for us to use the t test when population variances are unknown.

We will use both the one-tailed and the two-tailed tests to show how the t test can be employed for both one-sample and two-sample tests of means.

11.6.1 One-Sample Tests of Means

If the population variance is unknown and the sample size is small, then we can use the t statistic defined in Eq. 10.16 to test the null hypothesis associated with both one-tailed and two-tailed cases:

$$t_v = \frac{\overline{X} - \mu}{s_X / \sqrt{n}} \tag{10.16}$$

Example 11.3 Average Mileage of a Moving Van. United Van Lines Company is considering purchasing a large, new moving van. The sales agency agreed to lease the truck to United Van Lines for 4 weeks (24 working days) on a trial basis. The main concern of United Van Lines is the miles per gallon (mpg) of gasoline that the van obtains on a typical moving day. The mpg values for the 24 trial days were

8.5	9.5	8.7	8.9	9.1	10.1	12.0	11.5	10.5	9.6
8.7	11.6	10.9	9.8	8.8	8.6	9.4	10.8	12.3	11.1
10.2	9.7	9.8	8.1						

United Van Lines will purchase the van if it is convinced that the average value for miles per gallon is greater than 9.5.

To perform the hypothesis testing, we define the null and alternative hypothesis tests as

$$H_0 : \mu \leq 9.5$$
$$H_1 : \mu > 9.5$$

Table 11.2 Values of t_α versus z_α

	t_α value			
Sample size	*.10*	*.05*	*.025*	*.01*
10	1.372	1.812	2.228	2.764
20	1.325	1.725	2.086	2.528
120	1.289	1.658	1.980	2.358
∞	1.282	1.645	1.960	2.326
z_α	1.282	1.645	1.960	2.326

```
MTB > SET INTO C1
DATA> 8.5 9.5 8.7 8.9 9.1 10.1 12.0 11.5 10.5 9.6 8.7 11.6 10.9 9.8 8.8 8.6 9.4
DATA> 10.8 12.3 11.1 10.2 9.7 9.8 8.1
DATA> END
MTB > TINTERVAL WITH 95% CONFIDENCE USING C1

Confidence Intervals

Variable    N     Mean    StDev     SE Mean        95.0 % CI
C1         24    9.925    1.189       0.243     (9.423,  10.427)
MTB > TTEST 9.5 C1

T-Test of the Mean

Test of mu = 9.500 vs mu not = 9.500
Variable    N     Mean    StDev     SE Mean          T        P
C1         24    9.925    1.189       0.243       1.75    0.093
```

Fig. 11.13 MINITAB output for Example 11.3

The significance level for this test is $\alpha = .05$.

The MINITAB output for Example 11.3 is presented in Fig. 11.13. From this output, we calculate the test statistic as

$$t = \frac{\overline{X} - 9.5}{s_X/\sqrt{n}} = \frac{9.925 - 9.5}{.243} = 1.75$$

From Table A4, we find that $t_{.05,23} = 1.714$. Because $1.75 > 1.714$, we reject H_0—and advise United Van Lines to buy the van.

11.6.2 Two-Sample Tests of Means

To test the difference between two means when the population variances are unknown and the samples are small, we use the t statistic of Eq. 11.17, which is similar to Eq. 10.16. Here, two populations are normally distributed, and the two samples that are used to do the test are independent of each other. The hypotheses for a two-tailed case can be expressed as

$$H_0 : \mu_1 - \mu_2 = D$$
$$H_1 : \mu_1 - \mu_2 \neq D$$

where D can be either zero or a positive number. The statistic for testing the hypotheses can be defined as

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - D}{s\sqrt{\frac{n_1+n_2}{n_1 n_2}}} \tag{11.17}$$

This statistic has a t distribution with (n_1+n_2-2) degrees of freedom and where

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \tag{11.18}$$

is the pooled variance. Note that the t statistic denned in Eq. 11.17 also can be used in a one-tailed test. Two examples are used to show how Eq. 11.17 can be used to do both two-tailed and one-tailed tests.

Example 11.4 Competitive Versus Coordinative Bargaining Strategies. We now use a real-world example to show how the t statistics defined in Eqs. 11.17 and 11.18 can be used to test whether the competitive bargaining strategy differs in its results from the coordinative bargaining strategy. This example is adapted from S. W. Clopton's research (*Journal of Marketing Research,* February 1984) in which he compared so-called competitive and coordinative bargaining strategies in buyer–seller negotiations. Inflexibility in an effort to force concessions best defines competitive bargaining. A coordinative bargaining strategy, however, involves a great deal more cooperation and more of a problem-solving approach.

One of Clopton's negotiation experiments involved a sample of 16 organizational buyers. Clopton reported that in negotiations in which the maximum profit was fixed, the sample participants were perfectly divided in their choice of strategy; that is, 8 buyers employed a competitive bargaining strategy and the other 8 buyers used a coordinative approach.

The table lists the individual savings for the two groups of buyers. Using $\alpha = .05$, test to find if there is a difference in mean buyer savings for the two strategies.

	Competitive	Coordinative
Sample size	8	8
Mean savings	$1,706.25	$2,106.25
Standard deviation	$ 532.81	$ 359.99

$$H_0 : (\mu_1 - \mu_2) = 0$$
$$H_1 : (\mu_1 - \mu_2) \neq 0$$

Since sample size is only 8 for each, the t test statistic of Eq. 11.17 should be used to do the test. Substituting related information into Eqs. 11.18 and 11.17, we obtain

$$\begin{aligned} s^2 &= \frac{(8-1)(532.81)^2 + (8-1)(359.99)^2}{8+8-2} \\ &= 206,739.648 \\ t &= \frac{1,706.25 - 2,106.25}{\sqrt{206,739.648\left(\frac{1}{8}+\frac{1}{8}\right)}} \\ &= \frac{-400}{227.343} = -1.759 \end{aligned}$$

From Table A4 in Appendix A, we find $t_{14,\ .025} = 2.145$. Because 1.759 is smaller than 2.145, the null hypothesis cannot be rejected.

Example 11.5 The Effect of a Moderator on the Number of Ideas Generated. Fern (1982) studied the impact of the presence of a moderator on the number of ideas generated by groups.[4] He first randomly sampled 4 groups that included a moderator. Then he independently and randomly sampled another four groups that lacked a moderator. The mean number of ideas generated and the sample standard deviation for the two sets of samples were:

First set of samples: $\bar{X}_1 = 78.00$, $s_1 = 24.4$, $n_1 = 4$
Second set of samples: $\bar{X}_2 = 63.5$, $s_2 = 20.2$, $n_2 = 4$

Let μ_1 and μ_2 represent the respective population means for groups with and without a moderator. Then the test can be defined as

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 > 0$$

The significance level for this test is $\alpha = .05$. To perform the test, we substitute all related information into Eqs. 11.17 and 11.18 and obtain

$$\begin{aligned} s^2 &= \frac{(3)(24.4)^2 + (3)(20.2)^2}{4+4-2} = 501.7 \\ s &= \sqrt{501.7} = 22.4 \end{aligned}$$

[4] Fern E.F.: The use of focus groups for idea generators: The effect of group size, acquaintanceship, and moderator on response quantity and quality. J. Mark. Res. **19**, 1–13 (1982)

Then,

$$t_6 = \frac{78.0 - 63.5}{22.4\sqrt{\frac{8}{16}}} = .915$$

From Table A4 in Appendix A of this book, we find $t_{6,0.05} = 1.943$. Because .915 is smaller than 1.943, the null hypothesis of equality of population means cannot be rejected.

11.7 Hypothesis Testing for a Population Proportion

In Sect. 10.6, we discussed the confidence intervals for a population proportion. The Z statistic for the sample proportion $(\hat{P})$ and the confidence interval for the population proportion (P) are repeated here for convenience:

$$Z = \frac{\hat{P} - P}{\sqrt{\hat{P}(1-\hat{P})/n}} \tag{10.10b}$$

$$1 - \alpha = P\left\{\hat{P} - Z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} < P < \hat{P} + Z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right\} \tag{10.11}$$

where $z_{a/2}$ is the number such that $P(Z > z_{\alpha/2}) = \alpha/2$. The sample standard deviation used in Eq. 10.11 can be defined as

$$S_{\hat{P}} = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \tag{11.19}$$

Note that $\hat{P}$ instead of P is used to estimate $s_{\hat{P}}$ because P is unknown and must be replaced by its estimate, $\hat{P}$.

The procedure discussed in Sects. 11.4, 11.5, and 11.6 for testing population means for both one and two samples can be used to test the population proportion. Table 11.3 compares the null hypothesis for testing population means with that for testing population proportions.

In Table 11.3, both D and C can be zero or nonzero. If the sample size is large, the Z statistic should be used to do the null hypothesis test; if the sample size is small, the t statistic should be used. The Z statistic for testing one population proportion is

Table 11.3 Null hypothesis for testing population means and population proportions

	Population means		Population proportions	
	One sample	*Two samples*	*One sample*	*Two samples*
1.Upper-tailed test	H_0: $\mu = D$	$\mu_1 - \mu_2 = D$	$P = C$	$P_1 - P_2 = C$
	H_1: $\mu > D$	$\mu_1 - \mu_2 > D$	$P > C$	$P_1 - P_2 > C$
2.Lower-tailed test	H_0: $\mu = D$	$\mu_1 - \mu_2 = D$	$P = C$	$P_1 - P_2 = C$
	H_1: $\mu < D$	$\mu_1 - \mu_2 < 0$	$P < C$	$P_1 - P_2 < C$
3.Two-tailed test	H_0: $\mu = D$	$\mu_1 - \mu_2 = D$	$P = C$	$P_1 - P_2 = C$
	H_1: $\mu \neq D$	$\mu_1 - \mu_2 \neq D$	$P \neq C$	$P_1 - P_2 \neq C$

$$Z = \frac{\hat{P} - P_0}{\sqrt{P_0(1 - P_0)/n}} \tag{11.20}$$

where P_0 is the value of P specified in H_0. Equation 11.20 is obtained by substituting P_0 for P in Eq. 10.19. The Z statistic for testing the difference between two population proportions is defined as

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}} \tag{11.21}$$

where $\bar{P}$ is defined as

$$\frac{\hat{P}_1 n_1 + \hat{P}_2 n_2}{n_1 + n_2}$$

Now let's see how the Z statistic of Eqs. 11.20 and 11.21 for testing proportions can be applied.

Example 11.6 The Promotability of Company Employees. Francis Company is evaluating the promotability of its employees—that is, determining the proportion of employees whose ability, training, and supervisory experience qualify them for promotion to the next level of management. The human resources director of Francis Company tells the president that 80 percent of the employees in the company are "promotable." However, a special committee appointed by the president finds that only 75 percent of the 200 employees who have been interviewed are qualified for promotion. Use this information to do a two-tailed null hypothesis test at $\alpha = 5\ \%$:

$$H_0 : P = .80$$
$$H_1 : P \neq .80$$

From Table A3, we know that we should reject H_0 if $Z > z_{.025} = 1.96$ or if $Z < -z_{.025} = -1.96$.

Table 11.4 Sample data on canned food from old and new plants

Plant	Mean defect rate from each lot	Size of sample
New	$\hat{P}_1 = .065$	$n_1 = 50$
Old	$\hat{P}_2 = .052$	$n_2 = 40$

Substituting $p = .75$, $p_0 = .80$, and $n = 200$ into Eq. 11.20, we obtain

$$Z = \frac{.75 - .80}{\sqrt{\frac{(.8)(1-.8)}{200}}} = \frac{-.05}{.0283}$$
$$= -1.7668$$

Because $-1.7668 > -1.96$, we cannot reject H_0. In other words, the percentage of "promotable" employees is 80 %.

Example 11.7 Defects in Canned Food. A food manufacturer has two canning plants. The company's management wants to know whether the mean defect rate of a canned food from the new plant is different than that of the same canned food from the old plant. The canned food is packed in a carton that holds 24 cans. There are 500 cartons in each lot. Table 11.4 gives the sample data obtained from each plant.

The hypotheses to be tested in terms of Eq. 11.21 are

$$H_0 : P_1 - P_2 = 0$$
$$H_0 : P_1 - P_2 \neq 0$$

First we calculate $\hat{P}_1 - \hat{P}_2$ and the standard derivation of $(\hat{P}_1 - \hat{P}_2)$ as follows: $\hat{P}_1 - \hat{P}_2 = .065 - .052 = .013$

$$\overline{P} = \frac{\hat{P}_1 n_1 + \hat{P}_1 n_2}{n_1 + n_2} = \frac{(.065)(50) + (.052)(40)}{50 + 40} = .059$$
$$s = \sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}$$
$$= \sqrt{\frac{(.059)(.941)}{50} + \frac{(.059)(.941)}{40}} = .05$$

where s = standard deviation of $(\hat{P}_1 - \hat{P}_2)$. If we specify $\alpha = .05$ and $z_{.025} = 1.96$, then using the Z-value approach, we have

$$Z = \frac{.013}{.05} = .26$$

Z is smaller than 1.96, so we cannot reject H_0. In other words, the management confirms that the mean defect rate of the new plant is not statistically different from the mean defect rate of the old plant.

11.8 Chi-Square Tests of the Variance of a Normal Distribution

In Chap. 10, we discussed confidence intervals for the variance. Now it is time to consider how to conduct hypothesis tests on the variance from a normal population. When we conducted tests on the population mean μ_x, we based our test on the sample mean $\bar{X}$. Thus, it seems natural that when we conduct tests of the population variance σ_X^2, we base our tests on the sample variance s_X^2. From Chaps. 9 and 10, we know that

$$\chi_{n-1}^2 = \frac{(n-1)s_X^2}{\sigma_X^2}$$

which follows a chi-square distribution with $(n-1)$ degrees of freedom. We are interested in testing whether the population variance is equal to some specific value, σ_X^{*2}; that is,

$$H_0 : \sigma_X^2 = \sigma_X^{*2}$$

Thus, when the null hypothesis is true, the random variable defined in Eq. 11.22 follows a chi-square distribution with $(n-1)$ degrees of freedom.

$$\chi_{n-1}^{*2} = \frac{(n-1)s_X^2}{\sigma_X^{*2}} \tag{11.22}$$

For many applications, we are concerned that the variance of our population may be equal to, larger than, or smaller than some specified value, σ_x^{*2}. The hypothesis testing on σ_x^2 can be defined as follows:

1. Two-tailed test

$$H_0 : \sigma_X^2 = \sigma_X^{*2}$$
$$H_1 : \sigma_X^2 \neq \sigma_X^{*2}$$
$$\text{Test statistics } \chi^2 = \frac{(n-1)s_X^2}{\sigma_X^{*2}}$$
$$\text{Reject } H_0 \text{ if } \chi^2 > \chi_{\alpha/2,n-1}^2 \text{ or if } \chi^2 < \chi_{1-\alpha/2,n-1}^2.$$

2. One-tailed test

$$\begin{array}{ll} H_0 : \sigma_X^2 \leq \sigma_X^* & H_0 : \sigma_X^2 \geq \sigma_X^* \\ H_1 : \sigma_X^2 > \sigma_X^* & H_1 : \sigma_X^2 < \sigma_X^* \end{array}$$

$$\text{Reject } H_0 \text{ if } \chi^2 > \chi^2_{\alpha,n-1}. \text{ Reject } H_0 \text{ if } \chi^2 < \chi^2_{1-\alpha,n-1}.$$

Example 11.8 Variability in Customer Waiting Time. Suppose the manager of a bank is thinking of introducing a "single-line" policy that directs all customers to enter a single waiting line in the order of their arrival and "feeds" them to different tellers as the latter become available. Although such a policy does not change the average time customers must wait, the manager prefers it because it decreases waiting-time variability. The manager's critics, however, claim that this variability will be at least as great as for a policy of multiple: independent lines {which in the past had a standard deviation of $\sigma_X^* =$ 6 min per customer. All have agreed to use a hypothesis test at the 5 % significance level to settle the issue. This test is to be based on the experience of a random sample of 20 customers on whom the new policy is "tried out." The two opposing hypotheses are

$$H_0 : \sigma_X^2 \geq 36$$
$$H_1 : \sigma_X^2 < 36$$

Here, 36 is chosen as the H_0 value even though any number greater than 36 is in H_0. The bank's statistician selects the test statistic as

$$\chi^2_{n-1} = \frac{(n-1)s_X^2}{\sigma_X^{*2}}$$

For a desired significance level of $\alpha = .05$ and 19 degree of freedom, Table A5 in Appendix A suggests a critical value of 10.117 (this being a lower-tailed test). Thus, the decision rule must be as follows: fail to reject H_0 if $\sigma^2_{1-0.05,19} = 10.117$. After taking a sample of 20 customers, the statistician finds the sample single-line waiting times to have a standard deviation of $s_X = 4$ min per customer. Accordingly, the computed value of the test statistic is

$$\chi^2_{n-1} = \frac{(n-1)s_X^2}{\sigma_X^{*2}} = \frac{4^2(20-1)}{36} = 8.44$$

Because 8.44 is smaller than 10.117, the null hypothesis should be rejected at the 5 % significance level, which means that the sample result is statistically significant. In other words, the observed divergence from the hypothesized value of $\sigma_X^* = 6$ min is not likely to be the result of chance factors operating during sampling.

11.9 Comparing the Variances of Two Normal Populations

In the last section, we showed that an x^2 distribution can be used to test whether the population variance of a normal distribution is equal to a specific value. In Chap. 9, we showed that the ratio of two independent x^2 variables (each divided by its degrees of freedom) is an F random variable. The F random variable is defined as

$$F = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} \tag{11.23}$$

The F random variable follows an F distribution. If $\sigma_X = \sigma_Y$, then Eq. 11.23 reduces to $F = s_X^2/s_Y^2$. The F distribution has degrees of freedom (n_X–l) and (n_Y–l).

If we want to test whether σ_X^2 is equal to σ_y^2, the hypotheses are

$$H_0 : \sigma_X^2 = \sigma_Y^2$$
$$H_1 : \sigma_X^2 \neq \sigma_Y^2$$

Using the data of Example 11.5, we define the F statistic as $F = (24.4)^2/(20.2)^2 = 1.46$. The degrees of freedom are $(n_x–1) = 3$ and $(n_y–1) = 3$ (from Table A6 in Appendix A), so we have $F_{3,3,\ 0.05} = 9.28$. Because the alternative hypothesis is two-sided, this is the appropriate critical value for testing at the 10 % significance level. Clearly, 1.46 is much smaller than 9.28; the null hypothesis cannot be rejected. There is no evidence that variances are different in the two testing groups.

11.10 Business Applications

Application 11.1 EPS and Rates of Return for JNJ Versus Those for MRK. In Chap. 7, we calculated descriptive statistics for the earnings per share (EPS) and rates of return (R) for JNJ and MRK. This information is presented in Table 11.5. Using 20 years of EPS data (1990–2009), we can test whether JNJ's average EPS ($3.0375) is significantly different from MRK's EPS ($3.0898). Our sample consists of only 20 years of data, so we should use the t test instead of the Z test. The hypotheses to be tested are

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_l : \mu_1 - \mu_2 \neq 0$$

where μ_1 and μ_2 are the average EPS for JNJ and MRK, respectively.

Table 11.5 Annual EPS and returns for JNJ and MRK

	EPS		Return	
	JNJ	*MRK*	*JNJ*	*MRK*
Mean	3.0375	3.0898	1.7806	1.3545
Std. Dev.	0.8694	1.1640	0.3010	0.2124
Skewness	0.1257	0.8487	0.5253	0.4382

Because the sample is small, we use the pooled variance of Eq. 11.18 and the test statistic of Eq. 11.17. Substituting into Eqs. 11.17 and 11.18, we get[5]

$$t = \frac{3.0898 - 3.0375}{\sqrt{\frac{(20-1)(1.1640)^2+(20-1)(0.8694)^2}{20+20-2}\left[\frac{20+20}{20\times 20}\right]}} = \frac{0.0523}{0.32487} = 0.16099$$

From Table A4, we can see that the critical t value for $\alpha = .05$ with 40 degree of freedom is 2.021. Our test statistic has $n_1+n_2-2 = 20+20-2 = 38$ degree of freedom, so we compare 0.16099 to 2.021 and learn that we are unable to reject the null hypothesis that the average EPS of JNJ and MRK are the same.

Alternatively, we input EPS data into MINITAB and obtain the means, the standard deviations, the t statistic, and the p-value as follows:

Data Display
EPS(JNJ)

3.38	4.30	1.54	2.71	3.08	3.65	2.12	2.41	2.23	2.94	3.39
1.83	2.16	2.39	2.83	3.46	3.73	3.63	4.57	4.40		

Data Display
EPS(MRK)

4.51	5.39	1.70	1.86	2.35	2.63	3.12	3.74	4.30	2.45	2.90
3.14	3.14	3.03	2.61	2.10	2.03	1.49	3.64	5.68		

Two-Sample *T*-Test and CI: EPS(MRK), EPS(JNJ)
Two-sample T for EPS(MRK) vs EPS(JNJ)

N	Mean	StDev	SE	Mean
EPS(MRK)	20	3.09	1.16	0.26
EPS(JNJ)	20	3.038	0.869	0.19

[5] If the nonpooled variance is used, then the t value is

$$t = (3.0898 - 3.0375) \Big/ \sqrt{\frac{(1.1640)^2}{20} + \frac{(0.8694)^2}{20}} = \frac{0.0523}{0.32487} = 0.16099$$

Difference = mu (EPS(MRK)) - mu (EPS(JNJ))
Estimate for difference: 0.053
95 % CI for difference: (−0.607, 0.713)
T-Test of difference = 0 (vs not =): T-Value = 0.16 P-Value = 0.871 DF = 38

From this computer output, we find that the t statistic equals .16 and the p-value equals .871. Again, we are unable to reject the null hypothesis that the average EPS of JNJ and MRK are the same.

Application 11.2 Analysis of the Bank Risk Premium. The international banking crisis of 1974, involving the failure of the Franklin National Bank in New York, led the Federal Reserve System to guarantee the international as well as the domestic deposits of the bank.[6] Giddy (1980) hypothesized that this "Franklin Message" would lead to a decrease in the risk premium attached to large American banks' deposits. (Risk premium here is taken to be measured by the excess of secondary-market certificate of deposit rates over Treasury bill yields.) For 48 months before the "Franklin Message," the mean risk premium was .899 and the variance was .247. For 48 months after the message, the mean and variance were .703 and .320. If μ_1 and μ_2 denote the means before and after the message, respectively, test the null hypothesis

$$\mathrm{H}_0 : \mu_1 - \mu_2 = 0$$

against the alternative hypothesis

$$\mathrm{H}_1 : \mu_1 - \mu_2 > 0$$

Assume that the data can be regarded as independent random samples from the two populations.

The decision rule is to reject H_0 in favor of H_1 if

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > Z_\alpha$$

In this example, $\bar{X}_1 = .899, s_1^2 = .247, n_1 = 48, \bar{X}_2 = .703, s_2^2 = .320$, and $n_2 = 48$, so

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{.899 - .703}{\sqrt{\frac{.247}{48} + \frac{.320}{48}}} = 1.80$$

[6] This example is drawn from a study by Giddy I.H.: Moral Hazard and Central Bank Rescues in an international context. Financ. Rev. **15**(2), 50–60 (1980). Reprinted by permission of the publisher.

From Table A3 in Appendix A, we find that the value of α corresponding to $z_\alpha = 1.80$ is .0359. Hence, the null hypothesis can be rejected at all levels of significance greater than 3.59 %. If the null hypothesis of equality of population means were true, the probability of observing a sample result as extreme as or more extreme than that found would be .0359. This is quite strong evidence against the null hypothesis of equality of these means, suggesting rather a decrease in the mean risk premium after the "Franklin Message."

Application 11.3 Analysis of Rates of Return for Retail Firms.[7] In their study aimed at finding early warning signals of business failure, Sharma and Mahajan (1980) used a random sample of 23 failed retail firms that 3 years before showed a mean return on assets of .058 and a sample standard deviation .055. An independent random sample of 23 nonfailed retail firms showed a mean return of .146 and a standard deviation of .058 for the same period. If μ_1 and μ_2 denote the population means for failed and nonfailed firms, respectively, test the null hypothesis

$$H_0 : \mu_1 - \mu_2 \geq 0$$

against the alternative hypothesis

$$H_1 : \mu_1 - \mu_2 < 0$$

Assume that the two population distributions are normal and have the same variance.

The decision rule is to reject H_0 in favor of H_1 if

$$\frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{n_1+n_2}{n_1 n_2}}} < -t_{v,\alpha}$$

For these data, we have $\bar{X}_1 = .058, s_1 = .055, n_1 = 23, \bar{X}_2 = .146, s_2 = .058$, and $n_2 = 23$. Hence,

$$\begin{aligned} s^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(22)(.055)^2 + (22)(.058)^2}{23 + 23 + -2} = .0031945 \end{aligned}$$

so that $s = \sqrt{.0031945} = .0565$. Then,

[7] This example is taken from Sharma S., Mahajan V.: Early warning indicators of business failure. J. Mark. **44**, 80–89 (1980). Reprinted by permission of the American Marketing Association.

$$\frac{\overline{X}_1 - \overline{X}_2}{s\sqrt{\frac{n_1+n_2}{n_1 n_2}}} = \frac{.058 - .146}{.0565\sqrt{\frac{23+23}{23\times 23}}} = -5.282$$

For a 1 % level test, we have, by interpolation from Table A4, Student's t distribution with 44 (23 + 23−2) degrees of freedom, $t_{44,.01} = 2.414$. Because −5.282 is much less than −2.414, the null hypothesis is rejected at $\alpha = 1$ %. The data cast considerable doubt on the hypothesis that the population mean return on assets is at least as large for failed than for nonfailed retail firms.

The test just discussed and illustrated is based on the assumption that the two population variances are equal. It is also possible to develop tests that are valid when this assumption does not hold.

Application 11.4 Hypothesis Testing Approach to Interpret the Quality Control Chart. To use a quality control chart as discussed in Sect. 10.9 is to perform a statistical test of a hypothesis each time a sample is taken and plotted on the chart. In general, the null hypothesis H_0 is that the process is in control, and the alternative hypothesis H_1 is that the process is out of control. For example, in an $\bar{X}$ -chart, to determine whether the process mean has shifted, we can test the hypothesis

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

where μ is the population mean and μ_0 is the specified value for μ. This null hypothesis can be converted into a confidence interval at a specified α value, as we noted in Sect. 11.5. This confidence interval defines the upper and lower limits of a control chart.

The $\bar{X}$ -chart for Consolidated Auto Supply Company given in Application 10.1 illustrates how a hypothesis testing approach can be used to interpret the quality control chart. Under the assumption of $\alpha = .27$ %, $UCL_X = 10.8210$, and $LCL_X = 10.6132$, it is found that 1 out of 24 sample means is smaller than 10.6132. Because $\frac{1}{24} = 4.17$ % is larger than .27 %, the null hypothesis should be rejected.

Application 11.5 Comparison of Organizational Values at Two Different Companies. Professor J. M. Liedtka used survey data from two firms, company A and company B.[8] Nine managers from each company were asked to rate the importance of each of a given list of organizational values on a scale of 1–7. Table 11.6 gives the ratings ascribed to these values by each company's managers. By using the t statistic as indicated in Eqs. 11.17 and 11.18, Liedtka performed a test. The t-values are indicated in the last column of Table 11.6. From this table, we see that the managers of company A rated industry leadership most important,

[8] Liedtka J.M.: Value congruence: The interplay of individual and organizational value systems. J. Bus. Ethics **8**, 805–815 (1989). Reprinted by permission of Kluwer Academic Publishers.

Table 11.6 Organizational values at companies A and B[a]

	Company A		Company B		
Value	*Score*	*Standard deviation*	*Score*	*Standard deviation*	*t*-value
Industry leadership	6.4	.5	5.6	1.0	−2.3[c]
Reputation of the firm	6.1	.8	5.9	.6	−.7
Employee welfare	5.0	1.0	3.0	.9	−4.5[d]
Tolerance for diversity	5.0	.9	3.9	1.4	−2.1[b]
Service to the general public	3.6	1.9	3.7	.9	.2
Value to the community	3.6	1.8	3.8	1.1	.3
Stability of the organization	5.3	1.2	3.3	1.3	−3.3[d]
Budget stability	4.3	1.5	4.6	1.3	.3
Organizational growth	5.2	1.6	4.9	1.1	−.5
Profit maximization	5.6	.7	6.7	.5	3.8[d]
Innovation	5.7	1.3	4.0	1.3	−2.9[c]
Honesty	5.9	1.1	4.7	1.8	−1.8[b]
Integrity	6.0	1.1	4.4	2.2	−1.9[b]
Product quality	6.0	.9	4.0	1.6	−3.3[d]
Customer service	5.0	1.3	4.0	1.9	−1.3
Average score	5.3	1.2	4.4	1.3	

Source: Adapted from Jeanne M. Liedtka (1989), "Value Congruence: The Interplay of Individual and Organizational Value Systems," *Journal of Business Ethics* 8. Reprinted by permission of Kluwer Academic Publishers.

[a]Score is based on a Likert Scale of 1 (of lesser importance) to 7 (of greater importance).

[b]Significant at alpha of. 10

[c]Significant at alpha of .05

[d]Significant at alpha of .01

followed by reputation of the firm integrity and product quality. The managers of company B rated profit maximization most important, followed by reputation of the firm and industry leadership. From the *t*-values presented in Table 11.6, it is evident that the ratings of organizational values differed significantly at $\alpha = .10$ for most of the items but not for budget stability, organizational growth, and customer service.

11.11 Summary

Using the concepts of statistical distributions and interval estimates, we showed how these concepts can be employed to test hypotheses about the parameters of a population. Hypothesis tests for one-tailed and two-tailed tests for both large and small samples were analyzed in detail. In addition to using the normal and *t* distributions for performing hypothesis testing, we discussed the use of the chi-square distribution to test null hypotheses about the sample variance from a normally distributed population.

The statistical concepts and methods discussed in the last 11 chapters will be used in the remaining 10 chapters to conduct further statistical analyses.

Questions and Problems

1. For each of the following, test the indicated hypothesis.

   (a) $n = 16$, $\bar{x} = 1{,}550$, $s^2 = 12$, H_0: $\mu = 1{,}500$, H_1: $\mu > 1{,}500$, $\alpha = .01$
   (b) $n = 9$, $\bar{x} = 10.1$, $s^2 = .81$, H_0: $\mu = 12$, H_1: $\mu \neq 12$, $\alpha = .05$
   (c) $n = 49$, $\bar{x} = 17$, $s = 1$, H_0: $\mu \geq 18$, H_1: $\mu < 18$, $\alpha = .05$

2. The estimated variance based on 4 measurements of a spring tension was .25 g. The mean was 37 g. Test the hypothesis that the true value is 35 g. Use $\alpha = .10$ and H_1: $\mu > 35$.
3. A population has a variance σ^2 of 100. A sample of 25 from this population had a mean equal to 17. Can we reject H_0: $\mu = 21$ in favor of H_1: $\mu \neq 21$? Let $\alpha = .05$.
4. Suppose a sample of 15 rulers from a given supplier has an average length of 12.04 in. and the sample standard deviation is .015 in.. If α is .02, can we conclude that the average length of the rulers produced by this supplier is 12 in., or should we accept H_1: $\mu \neq 12.00$?
5. The drained weights, in ounces, for a sample of 15 cans of fruit are given below. At a 5 % level of significance, use MINITAB to test the hypothesis that on average a 12-oz drained-weight standard is being maintained. Use H_1: $\mu \neq 12.0$ as the alternative hypothesis.

12.0	12.1	12.3	12.1	12.2
11.8	12.1	11.9	11.8	12.1
12.4	11.9	12.3	12.4	11.9

6. An advertisement for a brand-name camera stated that the cameras are inspected and that "60 % are rejected for the slightest imperfections." To test this assertion, you observe the inspection of a random selection of 30 cameras and find that 15 are rejected. Construct a test, using $\alpha = .05$.
7. A 1984 study indicated that the average yearly housing cost for a family of 4 was $12,983. A random sample of 200 families in a US city resulted in a mean of $ 14,039 with a standard deviation of $2,129. Is this city's sample mean significantly higher than the population mean? Use α–.05.
8. The data entry operation in a large computer department claims that it gives its customers a turnaround time of 6.0 h or less. To test this claim, one of the customers took a sample of 36 jobs and found that the sample mean turnaround time was $\bar{x} = 6.5$ h with a sample standard deviation of $s = 1.5$ h. Use H_0: $\mu = 6.0$, H_1: $\mu > 6.0$, and $\alpha = .10$ to test the data entry operation's claim.
9. The following data represent the time, in seconds, that it took the sand in a sample of timers to run out. At the 10 % significance level, can we conclude that the mean for timers of this type is not equal to the nominal 3 min?

190	199	198	176	180	174
181	183	208	188	198	165

(a) Use H_1: $\mu \neq 180$ as the alternative hypothesis.
(b) Use MINITAB to test (1) H_1: $\mu \neq 180$ and (2) H_1:$\mu > 180$.

10. Independent random samples from normal populations with the same variance gave the results shown in the following table. Can we conclude that the difference between the means, μ_1–μ_2, is less than 5? That is, test H_0: μ_1–$\mu_2 \geq 5$ with $\alpha = .05$.

Sample	n	Mean	Standard deviation
1	15	22	9
2	9	25	7

11. What is hypothesis testing? Why are we interested in hypothesis testing? In hypothesis testing, is it possible to prove a hypothesis true?
12. What are the types of errors that can be made in hypothesis testing? Which type of error is generally regarded as more serious?
13. For each of the following pairs of hypotheses, explain what the null hypothesis should be.

(a) Not guilty versus guilty in a court case.
(b) Cage is safe versus cage is unsafe when testing the safety of lion cages.
(c) New drug is safe to use versus new drug is unsafe when determining whether the FDA should allow a new arthritis medicine to be sold.
(d) New treatment is safe versus new treatment is unsafe when determining whether the FDA should allow a new treatment for AIDS to be used.

14. Compare the concepts of interval estimation discussed in Chap. 10 with the concept of hypothesis testing discussed in this chapter. How are they related?
15. Compare a one-tailed test with a two-tailed test. Give some examples wherein a one-tailed test is preferable to a two-tailed test. Give some examples wherein a two-tailed test is preferable to a one-tailed test.
16. Briefly explain what is meant by the power of a test. Why is the power of the test important?
17. What is a simple hypothesis? What is a composite hypothesis? Give some examples of a simple hypothesis. Give some examples of a composite hypothesis.
18. In 1981, the election for governor of the state of New Jersey in which Tom Kean defeated Jim Florio was so close that Florio demanded a recounting of the votes. If you were Florio and you were conducting a hypothesis test of who won the election, what would your null hypothesis be? How would your answer change if you were Kean?
19. In conducting a hypothesis test, how do we determine the rejection region?
20. Briefly explain why the central limit theorem is important in hypothesis testing.
21. Evaluate the following statement: "If we reject the null hypothesis that $\mu = \mu_0$ in a two-tailed test, we will also reject it in a one-tailed test (using the same α)."

22. Find the critical values for the following standard normal distributions:

    (a) Two-tailed test for $\alpha = .05$
    (b) One-tailed test for $\alpha = .05$
    (c) Two-tailed test for $\alpha = .01$
    (d) One-tailed test for $\alpha = .01$
    (e) Two-tailed test for $\alpha = .10$
    (f) One-tailed test for $\alpha = .10$

23. You are given the information $\bar{x} = 10$, $\sigma = 2$, and $n = 35$. Conduct the following hypothesis test at the .05 level of significance:

$$H_0 : \mu = 0 \text{ versus } H_1 : \mu > 0$$

24. Use the information given in question 23 to test

$$H_0 : \mu = 0 \text{ versus } H_1 : \mu \neq 0$$

at the .05 level of significance.

25. You are given the information $\bar{x} = 150$, $\sigma = 30$, and $n = 20$. Conduct the following hypothesis test at the .01 level of significance:
    H_0: $\mu = 100$ versus H_1: $\mu > 100$
26. Use the information given in question 25 to test
    H_0: $\mu = 100$ versus H_1: $\mu \neq 100$
    at the .01 level of significance.
27. You are given the information $\bar{x} = 1{,}050$, $s_x = 250$, and $n = 20$. Conduct the following hypothesis test at the .10 level of significance:
    H_0: $\mu = 1{,}100$ versus H_1: $\mu < 1{,}100$
28. Use the information given in question 27 to test
    H_0: $\mu = 1{,}100$ versus H_1: $\mu \neq 1{,}100$
    at the .10 level of significance.
29. A sample of 100 students in a high school has a sample mean score of 550 on the math portion of the SAT. Assuming that the sample standard deviation is 75, test, at the .05 level of significance, the hypothesis that the high school's mean SAT score is 500 against the alternative hypothesis that the school's mean SAT score does not equal 500.
30. Redo question 29, testing H_0: $\mu = 500$ against H_1: $\mu > 500$.
31. A sample of 20 students in a high school has a sample mean score of 520 on the English portion of the SAT. If the sample standard deviation is 65, test, at the .01 level of significance, the hypothesis that the school's mean SAT score is equal to 500 against the alternative hypothesis that the school's mean SAT score does not equal 500.
32. Redo question 31, substituting the alternative hypothesis H_1: $\mu > 500$.
33. Suppose a random sample of 25 people at a local weight-loss center is taken and the mean weight loss is found to be 12 lb. From past history, the standard

deviation is known to be 3 lb. Test the hypothesis that the mean weight loss for all the members of the weight-loss center is 10 lb against the alternative that it is more than 10 lb. Do the test at the .05 level of significance.

34. Redo question 33, but assume that the standard deviation is not known and that 3 lb represents the sample standard deviation. Do the test at the 5 % level of significance.
35. A quality control engineer is interested in testing the mean life of a new brand of light bulbs. A sample of 100 light bulbs is taken, and the sample mean life of these light bulbs is found to be 1,075 h. Suppose the standard deviation is known and is 100 h. Use a .05 level of significance to test the hypothesis that the mean life of the new bulbs is greater than 1,000 h.
36. Suppose that the quality control engineer in question 35 does not know what the standard deviation is and therefore uses the sample standard deviation. Does your answer to question 35 change? Why or why not?
37. Suppose that the quality control engineer in question 35 does not know what the standard deviation is and that this time, he selects a random sample of only 25 light bulbs. Does your answer to question 35 change? Explain.
38. An auditor is interested in the mean value of a company's accounts receivable. He randomly samples 200 accounts receivable and finds that the mean accounts receivable is $231. From past experience, he knows that the standard deviation is $25. Use a .01 level of significance to test whether the population mean accounts receivable is different from $200.
39. Use the information given in question 38 to test the hypothesis that the population mean accounts receivable is greater than $200 at the .05 level of significance.
40. An investment advisor is interested in determining whether a retirement community represents a potential clientele base. Of the 2,000 residents, he randomly samples 100 individuals and finds their mean wealth to be $525,000 with a sample standard deviation of $52,000. Use a .10 level of significance to test the hypothesis that the mean wealth is greater than $500,000.
41. An automobile manufacturer claims that a new car gets an average of 35 miles per gallon. Assume that the distribution is known to be normal with a standard deviation of 3.2 miles per gallon. A random sample of 10 cars gives an average of 35.1 miles per gallon. Test, at the .01 level of significance, the alternative hypothesis that the population mean is more than 35 miles per gallon.
42. Use the information given in question 41, except this time assume that the standard deviation is not known and that 3.2 miles per gallon represents the sample standard deviation. Again test, at the .01 level of significance, the alternative hypothesis that the population mean is at least 35 miles per gallon.
43. An aspirin manufacturer claims that its aspirin stops headaches in less than 30 min. A random sample of 100 people who use the pain killer finds that the average time it takes to stop a headache is 28.6 min with a sample standard

deviation of 4.2 min. Test, at the 5 % level of significance, the manufacturer's claim that this product stops headaches in less than 30 min.

44. Bob's SAT preparation service claims that the course it offers enables students to score an average of 600 or better on the math portion of the SAT. Suppose a random sample of 25 people taking the course has a mean score of 650 with a sample standard deviation of 50. Would it be more appropriate to use a one-tailed or a two-tailed test? Test the company's claim at the 10 % level of significance.
45. An advertising company claims that 80 % of stores that use their advertisements show increased sales. A random sample of 100 stores that used the company's advertisements reveals that 80 showed increased sales. Test, at the 5 % level of significance, whether at least 75 % of stores using the advertisements had increased sales.
46. Flip a coin 40 times and count the number of heads. Test, at the 5 % level of significance, whether the proportion of heads is .5.
47. A manufacturer claims that 95 % of its parts are free of defects. A random sample of 100 parts finds that 92 are free of defects. Test the manufacturer's claim at the 1 % level of significance.
48. An investment advisor claims that 70 % of the stocks she recommends will increase in price. Suppose testing a random sample of 125 stocks she recommends reveals that 75 have increased in price. Test her claim at the 10 % level of significance.
49. Ed's bar exam review claims that 90 % of the people who take its review course pass the bar exam on the first try. A random sample of 500 people who took the course reveals that 425 passed the bar exam on the first try. Test, at the 5 % level of significance, the null hypothesis that at least 90 % of those who take the review course pass the bar exam on the first try.
50. Use the information given in question 49 to test, at the 1 % level of significance, the null hypothesis that less than 80 % of those who take the course pass the bar exam on the first try.
51. In a taste test using 400 randomly selected people, 220 preferred a new brand of coffee to the leading brand. Test, at the 1 % significance level, the alternative hypothesis that at least 52 % prefer the new brand.
52. A popular commercial states that 4 out of 5 dentists who chew gum prefer sugarless gum. Suppose a random sample of 100 gum-chewing dentists is taken and 75 are found to prefer sugarless gum. Test, at the 10 % level of significance, the null hypothesis that the commercial's claim is true.
53. A diet center claims that people subscribing to its program lose an average of 4 lb in the first week of the diet. Suppose 25 people in the diet center's program are chosen at random and are found to have lost 4.3 lb in the first week with a sample standard deviation of 1.1 lb. Test, at the 5 % level of significance, the hypothesis that the mean weight loss is 4 lb.
54. Use the information given in question 53 and test, at the 5 % level of significance, the alternative hypothesis that the mean weight loss is at least 4 lb.

55. Suppose a farmer is interested in testing two fertilizers to see which is more effective. He uses the two fertilizers and gets the following results.

Fertilizer	Mean growth	Standard deviation	Size of sample
A	7 in.	.5 in.	100
B	6 in.	.2 in.	125

Test, at the 10 % level of significance, the hypothesis that the mean difference in growth between the two fertilizers is not significant.

56. A production manager is interested in the number of defects in batches derived from different production processes. He examines a random sample drawn from each process and records the following data:

Process	Mean defects	Standard deviation	Size of sample
A	221	25	90
B	300	80	110

Test, at the 1 % level of significance, the hypothesis that the mean difference in number of defects between the two production processes is not significant.

57. Suppose an attorney specializing in wage discrimination cases is interested in determining whether the earnings of men and women are significantly different. He collects the following data on earnings for a random sample of first-year accountants:

Sex	Mean earnings	Standard deviation	Size of sample
Female	$39,217	$12,210	125
Male	$43,121	$17,020	100

Test, at the 5 % level of significance, the hypothesis that the mean earnings of male and female first-year accountants do not differ significantly.

58. Suppose a political scientist is interested in whether wealth is a determining factor in the individual's propensity to vote. A random sample of 500 people who earned $ 100,000 or more showed that 390 voted, whereas a random sample of 400 people who earned less than $25,000 showed that 280 voted. Test, at the 10 % level of significance, the null hypothesis that the two population voting rates are equal against the alternative hypothesis that the voting rate is higher for people earning $ 100,000 or more.
59. A mutual fund manager claims that the returns of stocks in her fund have a variance of no more than .50. A random sample of 25 stocks in her fund has a sample variance of .72. Assuming that the distribution is normal, test the fund manager's claim at the 5 % level of significance.
60. Bob claims that the variance of the score for the people who took the SAT review course he offers is 100. Fred believes that Bob's students have a variance larger than 100. A random sample of 10 of Bob's students has a variance of 162. Test Fred's claim at the 10 % level of significance.

61. A political science professor believes students majoring in political science are more likely to vote in elections than students majoring in other disciplines. He collects the following information from two random samples of students:

Major	Proportion voting	Number of students
Political science	.65	120
Other	.62	113

Test, at the 5 % level of significance, this professor's hypothesis against a two-sided alternative that the population proportions are equal.

62. An education professor is interested in whether there is any difference between the proportion of students who have taken a review course that pass the bar exam and the proportion of those who have not taken a review course that pass the exam. She collects the following information from a random sample of students:

	Proportion passing	Number of students
Review course	.55	300
No review course	.49	400

Test the hypothesis that the population proportions are equal at the 10 % level of significance in terms of a two-tailed test.

63. Use the information given in question 62, but this time test the hypothesis that the proportion of students passing the exam is greater for those who take the course than for those who do not.
64. The IRS is interested in knowing whether people who have an accountant prepare their tax returns have fewer errors than people who prepare their own returns. A random sample of 500 people who had their returns professionally prepared reveals that 125 had errors. A random sample of 450 returns of people who prepared their own returns reveals that 128 had errors. Test, at the 1 % level of significance, the hypothesis that there is no difference between the number of errors for returns prepared by an accountant and the number of errors for returns prepared by the individual.
65. Use the information given in question 64 to test, at the 1 % level of significance, the hypothesis that the number of errors is greater for individuals who prepare their own returns than for people who have their returns professionally prepared.
66. A muffler manufacturer claims that the variance of its product is no more than 200. A random sample of 25 mufflers has a sample variance of 391. Assuming a normal distribution, test the manufacturer's claim at the 10 % level of significance.
67. An SAT review course claims that the variance of test scores of its graduates is less than 150. A random sample of 30 students who took the course is found to have a variance of 225. Assuming a normal distribution, test the review course's claim at the 10 % level of significance.

68. From past experience, a teacher finds that the variance of midterm test scores is 76. A random sample of 21 midterms in her course has a sample variance of 110. Assuming that the population is distributed normally, test whether the sample variance is different from the population variance at a 5 % level of significance.
69. Refer to question 4I to find the power of a 10 % level test when the true population mean mileage is 36 miles per gallon.
70. Referring to question 43, find the power of a 5 % level test when the true population mean time for headache relief is 35 min.
71. Assume that you're taking a part-time job in a zoo. You are called upon to inspect a new cage built to contain a ferocious lion. Do you set up the null hypothesis that the cage is safe or that the cage is dangerous?

Use the following information to answer questions 72–78. A college professor gives a test that has 10 true–false questions. Two students take the test. Student A, who does not know anything about the subject, answers the questions by tossing a coin. The college professor sets up the following hypothesis, where p represents the probability that a student gets an answer right.

H_0: The students do not know anything ($p = .5$).

H_1: The students know the subject ($p > .5$).

72. What is the consequence of a Type I error in this question?
73. What is the chance of student A getting exactly 6 correct answers when the null hypothesis is true?
74. If the professor decides to reject the null hypothesis (that means passing the student) when the students get eight or more correct answers, what is the probability of a Type I error?
75. If the professor wants to raise the standard for passing the test to nine or more correct answers, what is the probability of a Type I error?
76. Student B studies one night before the test, so he knows about 60 % of the material. What is the probability that this student can pass the test when the standard for passing is eight correct answers?
77. Plot the OC curve, assuming $p = .5, .6, .7, .8, .9, 1$.
78. Plot the power curve, assuming $p = .5, .6, .7, .8, .9, 1$.
79. A poll was done to predict the outcome of the upcoming election. Of the 900 potential voters who responded, 500 plan to vote for the incumbent. If a candidate needs 50 % of the votes to win the election, can you reject the hypothesis that the incumbent will win? Do a 5 % level of significance test.
80. On a given trading day, a financial economist randomly examines the stock prices of 500 companies and discovers that 205 went up and 295 went down. On this evidence, can he argue that more than 50 % of all the stocks went down in price? Do a 10 % level of significance test.
81. The head of the accounting department randomly examined some accounting entries and was upset with the high proportion of incorrect invoices. He instituted a new system to keep the proportion of bad invoices below 0.1 %. A year later, 10,000 invoices were randomly examined and six were found to be incorrect. Can this manager reject the null hypothesis that the proportion of bad invoices is 0.1 %? Do a 5 % level of significance test.

82. You are working for a consumer rights organization. You are interested in knowing whether the milk contained in 16-oz (1-pint) bottles really weighs 16 oz. You do not want to accuse the packer of cheating its customers unless you obtain convincing evidence. You collect 60 bottles of milk. The average weight is 15.32, and the standard deviation is 1 oz. Test at a 5 % significance level.
83. You are working for a VCR manufacturer. There are three shifts in the plant: morning shift, evening shift, and midnight shift. The manager suspects that the midnight shift's productivity is lower than 70 units. He wants to shut down the midnight shift without causing any labor-management tension. That means he will take that action only when he has enough evidence. Your responsibility is to test whether the productivity of the midnight shift is really lower than 70 units. You obtain the production for 100 nights and compute the mean as 68 and the standard deviation as 15. Test at a 5 % significance level. Propose your suggestion to the manager.
84. A college wants to increase its dormitory facilities to house 60 % of the students enrolled. In order to make sure that more than 60 % of the students want to live in the dormitory, the school randomly surveys 400 students and finds that 255 students intend to live in the dormitory. Can the school reject the null hypothesis of $p = 60$ %? Test at a 5 % significance level.
85. A cola company wants to change its formula for producing cola, but first it wants to make sure that more than 70 % of its customers will like the new cola better than the old. Two thousand people taste tested the cola, and 1,422 liked the new product better. Can the company reject the null hypothesis that only 70 % of its customers will like the new cola more? Do a 5 % test.
86. In order to control the job turnover ratio, the personnel department did a survey and found that out of the 500 employees who were hired in the last year, only 234 stayed. Does that provide enough evidence to support the hypothesis that the retention ratio is lower than 50 %? Do a 5 % test.
87. An insurance company wants to study the chances that a teenaged driver who owns a sports car will have an auto accident. Two thousand teenaged policyholders who own sports cars were sampled in the last year. Fifteen of them got into an accident and filed a claim for damages. Can the researcher reject the null hypothesis that less than 1 % of the policyholders got into accidents last year? Do a 5 % test.
88. A food company claims that its new product, low-fat yogurt, is 99 % fat-free. The management wants to keep the proportion of bad (not 99 % fat-free) products below 2 %. Inspectors check 500 cups of yogurt every month. In September, 20 cups of yogurt were discovered to be bad. Can you reject the null hypothesis that less than 2 % of the product is bad? Do a 5 % test.
89. A questionnaire was sent to 500 of a dry cleaner's customers to solicit their opinions about service received. Twenty-three customers were found to be unhappy with the service. On this evidence, can you reject the null hypothesis that more than 10 % of the customers are unhappy? Do a 5 % test.
90. The dean of the school of business wants the proportion of A grades given out by his faculty members to be around 10 %. He randomly surveys 2,000 students in

50 classes and finds that of the 2,000 grades given, 198 were A. Can he reject the null hypothesis that the proportion of A grades is about 10 %? Do a 10 % test.

91. The placement office in a college wants to know whether experience with personal computers is important in obtaining a job. The placement director randomly selects 600 job openings and finds that 313 jobs require computer experience. On this evidence, can he support the hypothesis that more than half of the jobs in the market today require computer experience? Do a 5 % test.
92. The head accountant in a large corporation conducted a survey last year to study the proportion of incorrect invoices. Of the 2,000 invoices sampled, 25 were incorrect. To lower the proportion, he instituted a new system. A year later, he wants to know whether the new system worked. He collects 3,000 invoices and obtains 30 incorrect invoices. Can he argue that his new system has successfully lowered the proportion of incorrect invoices? Do a 5 % test.
93. A new medicine was invented to treat hay fever, but the new drug was found to have unpleasant side effects. An experiment on 5,000 women and 4,000 men showed that 100 women and 60 men suffered side effects after they took the medicine. Does the evidence support the hypothesis that the drug causes side effects in more women than men? Test at the 5 % level.
94. A company believed its new toothpaste to have an effect in controlling tooth decay among children. It randomly selected a group of 400 children and gave them the new toothpaste. Another 300 children were randomly selected also and given another brand of toothpaste. It was found that 30 children using the new toothpaste and 25 children using the other brand suffered tooth decay. Can the manufacturer legitimately argue that the new toothpaste is more effective in controlling tooth decay? Do a 5 % test.
95. The PPP cola company wants to determine what age groups like its product. It surveyed 500 teenagers and 600 middle-aged people and found that 300 teenagers and 350 middle-aged people liked PPP cola. Can the company conclude that PPP cola is more popular among teenagers than among middle-aged people? Do a 5 % test.
96. Wood et al. (1979) studied the impact of comprehensive planning on the financial performance of banks. They used 4 random samples to perform their study. The sample size n, average annual percent return on owner's equity $\bar{x}$, and sample standard deviation s are presented in the table.

Average Annual Percent Return on Net Income

Classification	n	$\bar{x}\%$	s
Comprehensive formal planners	26	11.928	3.865
Partial formal planners	6	9.972	7.470
No formal planning system	9	4.936	4.466
Control group	20	2.098	10.834

Source: Wood, D.R., LaForge, R.L.: The impact of comprehensive planning on financial performance. *Acad. Manag. J.* **22**, 516–526 (1979). Reprinted by permission of the publisher

(a) Use the data in this table to construct a 90 % confidence interval for the difference between the mean of the "comprehensive formal planners" group and that of the "no formal planning system" group.
(b) Perform a hypothesis test at $\alpha = 10$ %.

97. Professor Preston et al. (1978) studied the effectiveness of bank premiums (stoneware, calculators) given as an inducement to open bank accounts[9]. They randomly selected a sample of 200 accounts each for "premium offered" and "no premium offered." They found that 79 % of the accounts opened when a premium was offered and 89 % of accounts opened when a premium was not offered were retained over a 6-month period. Use these data to test whether $P_x = 79$ % is statistically different from $P_y = 89$ %. Do a 5 % test.
98. Use the data in the table to answer the following questions by using MINITAB. Current ratios for GM and Ford

Year	Ford	GM
81	1.02	1.09
82	.84	1.13
83	1.05	1.40
84	1.11	1.36
85	1.10	1.09
86	1.18	1.17
87	1.24	1.56
88	1.00	1.00
89	.97	1.72
90	.93	1.37

(a) Test whether the current ratio is equal to 1 for both GM and Ford, respectively, at $\alpha = .05$.
(b) Construct a 95 % confidence interval of current ratio for both GM and Ford.
(c) Test whether there is a difference between current ratios of Ford and GM at $\alpha = .05$. (Assume that their variances are different.)
(d) Test whether there is a difference between current ratios of Ford and GM at $\alpha = .05$. (Assume that their variances are equal.)

99. The result of a random sample of before and after weights of 11 participants in a weight-loss program is shown in the table to the right. Using MINITAB, test whether the average weight loss is at least 16 lb at the 5 % significance level. Weights before and after a weight-loss program

[9] Preston R.H., Dwyer F.R., Rudelius W.: The effectiveness of bank premiums. J. Mark. **42**(3), 39–101 (1978)

Before	After
187	168
200	177
218	201
205	190
192	170
175	159
191	172
200	185
206	184
231	202
240	215

100. Refer to Table 10.3, in which the sizes of U-bolts by taking samples of 5 every hour over three shifts from a population with mean μ and standard deviation σ. In Table 10.3, the means of the 24 samples are also given. By considering the means of the 24 samples as a random sample from a population with mean μ and standard deviation $\sigma/\sqrt{24}$, test whether the population mean μ is significantly deviates from 10 in..
101. A manager claims that the standard deviation in their mean delivery time is less than 2.5 days. A sample of 25 customers is taken. The average delivery time in the sample was 4 days with a standard deviation of 1.2 days. Suppose the delivery times are normally distributed; at 95 % confidence, test the manager's claim.
102. A candidate believes that more than 30 % of the citizens will vote for him. A random sample of 250 citizens was taken and 101 of them vote for the candidate.

    (a) State the null and alternative hypotheses.
    (b) Using the critical value approach, test the hypotheses at the $\alpha = 1$ % level of significance.
    (c) Using the p-value approach, test the hypotheses at the $\alpha = 1$ % level of significance.

103. A poll on the preference of two presidential candidates A and B is shown below.

Candidate	Voters surveyed	Voters favoring this candidate
A	500	292
B	400	225

    At 99 % confidence, test to determine whether or not there is a significant difference between the preferences for the two candidates.
104. To test whether a bonus plan will improve the monthly sale volume in units, the monthly sale volumes of six salespersons before and after a bonus plan were recorded. At 99 % confidence, determine whether the bonus plan has increased sales significantly.

Monthly sales		
Salesperson	After	Before
1	94	90
2	82	84
3	90	84
4	76	70
5	79	80
6	85	80

Appendix 1: The Power of a Test, the Power Function, and the Operating-Characteristic Curve

The main purpose of this appendix is to discuss the power of a test and the power function. The related concepts of the power curve and the operating-characteristic curve (OC curve) are also discussed.

The Power of a Test and the Power Function

The *power of a test* is the probability of rejecting H_0 when it is false. The probability is equal to $(1–\beta)$, where β denotes the probability of Type II error. Other things being equal, the greater the power of the test, the better the test. The formula used to calculate $(1–\beta)$ can be derived as follows:

$$\begin{aligned}
\text{Power} &= 1-\beta \\
&= P(\text{null hypothesis rejected when it is false}) \\
&= P\left(\frac{\overline{X}-\mu_0}{\sigma_X/\sqrt{n}} > z_\alpha\right) \\
&= P\left(\overline{X} > \mu_0 + \frac{z_\alpha \sigma_X}{\sqrt{n}}\right) \\
&= P\left(\frac{\overline{X}-\mu_1}{\sigma_X/\sqrt{n}} > \frac{\mu_0-\mu_1}{\sigma_X/\sqrt{n}} + z_\alpha\right) \\
&= P\left(Z > \frac{\mu_0-\mu_1}{\sigma_X/\sqrt{n}} + z_\alpha\right) \qquad (11.24)
\end{aligned}$$

where μ_1 is the true population mean when the null hypothesis is false.

The functional relationship defined in Eq 11.24 is called the *power function*. This power function is derived from the assumption of an upper-tailed test. Similarly, the

power function in terms of a lower-tailed test can be denned as Eq. 11.25, and the power in terms of a two-tailed test can be denned as Eq. 11.26:

$$1-\beta = P\left[Z < \frac{\mu_0 - \mu_1}{\sigma_X/\sqrt{n}} - z_\alpha\right] \tag{11.25}$$

$$1-\beta = P\left[Z > \frac{\mu_0 - \mu_1}{\sigma_X/\sqrt{n}} + z_{\alpha/2}\right] + P\left[Z < \frac{\mu_0 - \mu_1}{\sigma_X/\sqrt{n}} - z_{\alpha/2}\right] \tag{11.26}$$

Example 11A.1 Power Function and Type II Error. From Example 11.1, investigating the average weight of a bag of cat food, we express the two hypotheses as

$$\begin{aligned} &H_0 : \mu = 60 \\ &H_1 : \mu > 60 \end{aligned}$$

Using the data for Example 11.1 in the text, we have $z_{.05} = 1.645$, $n = 100$, and $s_x = 5$. We calculate $(1–\beta)$ for $\mu_1 = 60$, 60.5, 61, 61.5, and 62 in accordance with Eq. 11.24:

1. $\mu_1 = 60 \quad 1-\beta = P\left(Z > \dfrac{60-60}{5/10} + 1.645\right)$
$= P(Z > 1.645)$
$= .05$

2. $\mu_1 = 60.5 \quad 1-\beta = P\left(Z > \dfrac{60-60.5}{5/10} + 1.645\right)$
$= P(Z > .645)$
$= .25945$

3. $\mu_1 = 61 \quad 1-\beta = P\left(Z > \dfrac{60-61}{5/10} + 1.645\right)$
$= P(Z > -.355)$
$= .6387$

4. $\mu_1 = 61.5 \quad 1-\beta = P\left(Z > \dfrac{60-61.5}{5/10} + 1.645\right)$
$= P(Z > -1.355)$
$= .9123$

5. $\mu_1 = 62 \quad 1-\beta = P\left(Z > \dfrac{60-62}{5/10} + 1.645\right)$
$= P(Z > -2.355)$
$= .99075$

Table 11.7 The relationship among μ_1, β_1, and $(1–\beta_1)$

μ_1	$1–\beta$	β
60.0	.05	.95
60.5	.25945	.74055
61.0	.6387	.3613
61.5	.9123	.0877
62.0	.99075	.00925

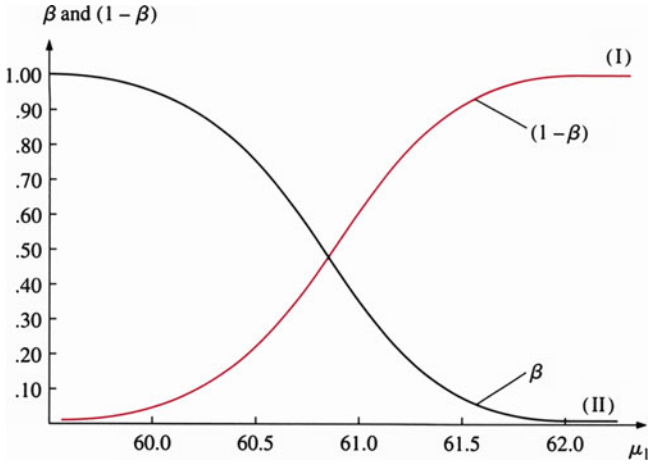


Fig. 11.14 The power function and the probability of Type II error

Using these results for $(1–\beta)$, we can easily calculate the values of β. Values for both $(1–\beta)$ and β are listed in Table 11.7. In Fig. 11.14, we present the relationship between the power function $(1–\beta)$ and the probability of Type II error (β). Curve I, which describes the relationship between $(1–\beta)$ and μ_1, is called the *power curve*. It is an increasing function of the value of μ_1. Curve II, which describes the relationship between β and μ_1, is called the *operating-characteristic* (OC) *curve*.

The OC curve is a decreasing function of μ_1. Overall, the relationship among μ_1, Type II error, and the power of the test can be summarized as follows: the larger the value of μ_1, the smaller the Type II error and the larger the power of the test. As we will see in the next section, the OC curve is frequently used in statistical quality control to analyze the risk involved in a sampling plan.

Operating-Characteristic Curve[10]

In Chap. 10, we constructed quality control charts by using sampling production process data. To collect sample data for quality control, we need sample plans that specify the lot size *N,* the sample size *n,* and the acceptance/rejection criterion.

[10] The material in this section draws heavily on Stevenson W.J.: Production/Operations Management, 3rd ed., pp. 829–836. Homewood, irwin), (1990)

An important feature of a sampling plan is how well it discriminates between lots of high quality and lots of low quality. The *operating-characteristic curve* can be used to describe the ability of a sample plan to differentiate high-quality lots from low-quality lots.

Acceptance sampling which has been discussed in Sect. 10.8 of Chap. 10 is frequently the most desirable method for quality control. The inspector takes a statistically determined random sample and applies a decision rule to determine the acceptance or rejection of the lot on the basis of the observed number of nonconforming items. The Type II error and Type I error discussed in this chapter are bases for the decision rule for statistical quality control. The probability that a lot containing the *lot tolerance percentage defective* (LTPD) will be accepted is known as the *consumer's risk,* or beta (β), or a Type II error. The probability that a lot containing the *acceptable quality level* (AQL) will be rejected is known as the *producer's risk,* or alpha (α), or a Type I error. Sampling plans are frequently designed so that they have a producer's risk of 5 % and a consumer's risk of 10 %, although other combinations also are used. Figure 11.15 shows an OC curve with the AQL, LTPD, producer's risk ($\alpha = 10$ %), and consumer's risk ($\beta = 10$ %). In Fig. 11.15, the horizontal axis represents lot quality (fraction defective) and the vertical axis represents probability of acceptance lot. We can see from the graph that a lot with 2 % defectives (the AQL) would have about 90 % probability of being accepted (and hence $\alpha = 1.00-.90 = .10$), Similarly, a lot with 17 % defective (the LTPD) would have about 10 % probability of being accepted (and hence $\beta = .10$).

It is possible to use trial and error to design a plan that will provide selected values for alpha (α) and beta (β) given the AQL and LTPD. In addition, Eq. 11.2 in the text can be used to determine the required sample size in a simple-sampling plan. However, standard references such as government MIU-STD (military standard) tables are widely used to obtain sample sizes and acceptance criteria for sample plans.

Example 11A.2 Construction of an OC Curve with Sample Size $n = 10$ and Defective Items $C = 1$. Suppose we want the OC curve for a situation in which a sample of $n = 10$ items is drawn from lots containing $N = 2{,}000$ items and lots are accepted if no more than $C = 1$ defective is found.

The ratio n/N 10/2,000 = .5 %, so it is small enough for us to use the binomial distribution.[11] In this case, $n = 10$ and $C = 1$. Table 11.8 presents the probability of acceptance (β) for 12 different fractions defective (P).

The β values indicated in Table 11.8 are obtained from Table Al in Appendix A. For example, when $n = 10$, $C = 1$, and P = .20, β is calculated as follows:

[11] Because the sampling is generally done without replacement, the hypergeometric distribution would be more appropriate if the ratio n/N exceeded 5 %.

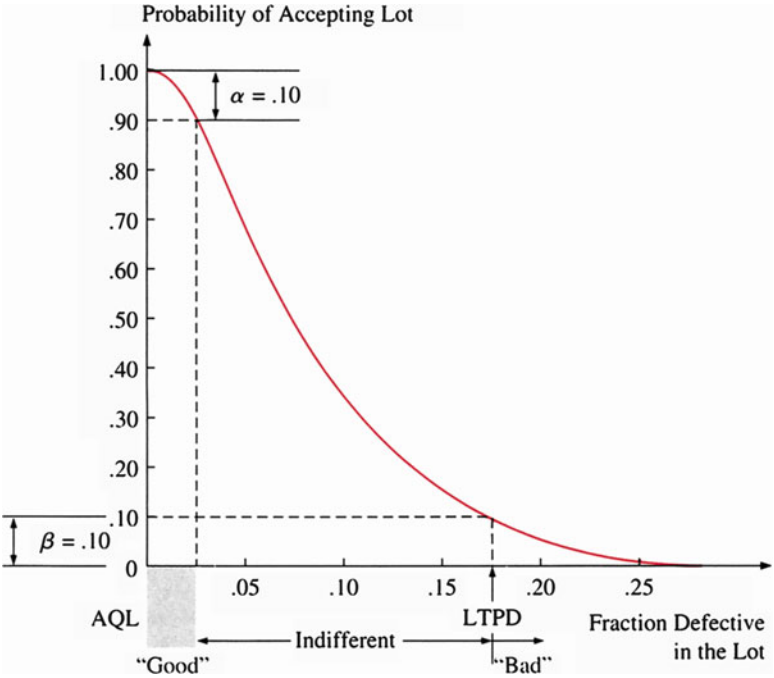


Fig. 11.15 The AQL indicates "good" lots; LTPD indicates "bad" lots (*Source:* W. J. Stevenson, *Production/Operations Management,* 3rd ed., 1990, p. 833, reprinted by permission of Richard D. Irwin)

Table 11.8 β for different fractions defective ($n = 10$, $C = 1$)

Fraction defective (P)	Probability of acceptance (β)
.05	.9139
.10	.7361
.15	.5443
.20	.3758
.25	.2440
.30	.1493
.35	.0860
.40	.0464
.45	.0233
.50	.0107
.55	.0045
.60	.0017

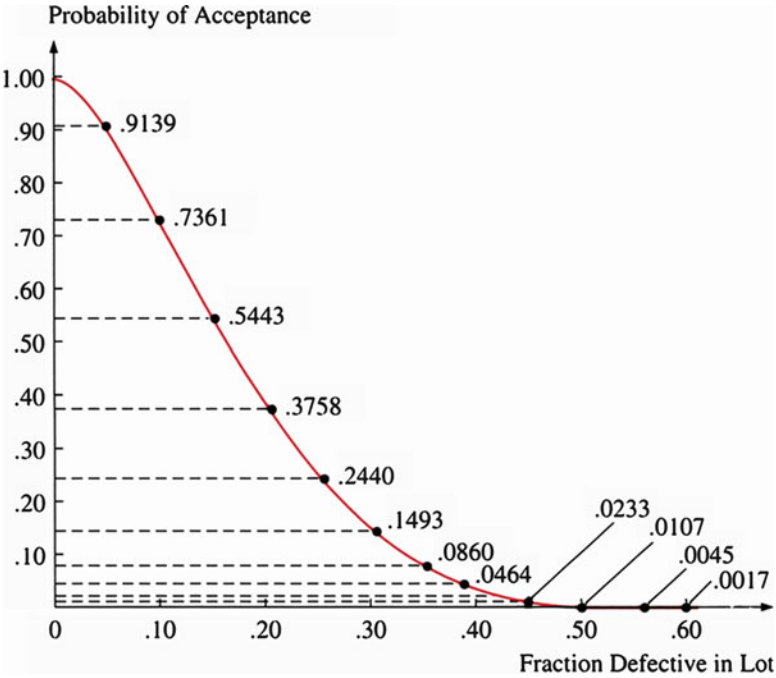


Fig. 11.16 OC curve for $n = 10$, $c = 1$ (*Source:* W. J. Stevenson, *Production/Operations Management,* 3rd ed., 1990, p. 835, reprinted by permission of Richard D. Irwin.)

$$\binom{10}{0}(.20)^0(1-.2)^{10} + \binom{10}{1}(.20)^1(1-.20)^{10-1} = .1074 + .2684 = .3758$$

By plotting all 12 different β values on a graph and connecting them, we get the OC curve shown in Fig. 11.16.

In theoretical statistics, only the power curve is generally considered. But in practical statistics, in certain types of problems, the OC curve is much easier to interpret for practical purposes such as quality control. Hence, the OC curve has been extensively used in quality control.

Up to this point, we have investigated the power function when the sample size is fixed. If sample size increases, then $\sigma_X/\sqrt{n}$ will decrease and the absolute value of $\frac{(\mu_0-\mu_1)}{\sigma_X/\sqrt{n}}$ will increase. From Eqs. 11.24, 11.25, and 11.26, it is obvious that the power $(1–\beta)$ increases. When n > 20 and p < .05, the Poisson distribution is useful in constructing OC curves for proportions. In effect, the Poisson distribution can be used to approximate the binomial distribution which has been discussed in Sect. 6.7 of Chap. 6.

A key aspect of an acceptance sampling plan is to offer protection, both to the consumer (who doesn't want to accept a bad lot) and to the producer/vendor (who doesn't want a good lot to be rejected). In this appendix, we have defined

α = producer's risk = probability of rejecting product of acceptable quality
β = consumer's risk = probability of accepting product of unacceptable quality

For a given sampling plan, the corresponding OC curve provides a value for the *consumer's risk* for each specified value of product quality (the proportion nonconforming). In Example 11A.2, the consumer will have a 37.58 % risk of accepting a lot containing 20 % nonconforming parts using that particular sampling plan. One option for the producer is to negotiate with the consumer for an acceptable quality level (AQL) that allows for a high chance of accepting a lot containing a low percentage of nonconforming parts.

It becomes clear that there is no one sampling plan that best fits all situations. Both the producer and the consumer want a plan that assures good quality yet preserves their individual interests. It is therefore appropriate to give careful consideration to the family of OC curves in the selection of a sampling plan.

Chapter 12
Analysis of Variance and Chi-Square Tests

Chapter Outline

Key Terms

Analysis of variance (ANOVA)	Between-treatments mean square
Factor	Within-treatment mean square
One-way ANOVA	Treatment effect
Two-way ANOVA	F distribution
Treatments	Scheffé's multiple comparison
Global (overall) mean	Blocking variable
Within-group variability	Goodness-of-fit tests
Between-groups variability	Chi-square test
Sum of squares	Observed frequency
Between-treatments sum of squares	Expected frequency
Within-treatment sum of squares	Contingency table

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_12,

12.1 Introduction

Both χ^2 and F distributions and their related testing statistics have been discussed in detail in the last three chapters. In this chapter, we will talk about how these two distributions can be used to do data analysis involving the means or the proportions of more than two populations. In other words, we will develop an understanding of (1) a technique known as *analysis of variance (ANOVA)*, which enables us to test the significance of the differences among sample means in terms of an F distribution and (2) tests of goodness of fit and independence in an χ^2 distribution. The ANOVA is used to test the equality of more than two population means. The goodness-of-fit test is used to test the equality of more than two population proportions or to assess the appropriateness of a distribution. The test of independence determines whether the differences among several sample proportions are significant or are instead likely to be due to chance alone.

First, we consider a one-way ANOVA model that has only one *factor* (characteristic) with several groups, such as different years of work experience or different types of tires. Then we explore both simple and simultaneous confidence intervals. Two-way analysis of variance with a single observation and more than one observation per cell is discussed in detail, as are tests of goodness of fit and independence. Finally, we consider applications of analysis of variance in business.

12.2 One-Way Analysis of Variance

In the analysis of variance, the F statistic is used to test whether the means of two or more groups are significantly different. It operates by breaking down the variance of the two or more populations into components. These components are then used to construct the sample statistic—hence, the term *analysis of variance*. The ANOVA can be used to analyze certain decisions, such as whether some products sell better when placed in certain sections of stores (e.g., as point-of-purchase, or impulse, sales), whether advertising is more effective in selling some products than in selling others, whether some employees are more motivated by some incentives than by others, and whether technology is variously effective in different workplaces. Furthermore, an accountant can use this technique to test whether the mean value of one set of sample accounts receivable is significantly different from another set or other sets.

The groups of data used to do the analysis of variance can be defined in terms of a single basis of classification (location, design, region, company, or the like) or by a dual classification. An ANOVA based on group data that are defined by a single classification is called *one-way ANOVA*. An ANOVA based on group data that are defined by a dual classification is called *two-way ANOVA*. In principle, both one-way and two-way analyses of variance are used to find out whether the means of all the populations considered are equal to one another.

12.2.1 Defining One-Way ANOVA

Suppose we want to test whether number of years of work experience since graduation has an effect on beginning salary for economics majors. The three *treatments* (or groups) are:

Treatment 1: Bachelor's degree with no work experience
Treatment 2: Bachelor's degree with 1 year of work experience
Treatment 3: Bachelor's degree with 2 years of work experience

We also assume that each student in this sample graduated from Rutgers University and specialized in labor economics. In order to simplify the necessary computations, we have restricted this to a random sample of only 12 observations—three samples (of four graduates) from each of the combinations. (A larger sample size would yield more convincing results.) Table 12.1 gives the 12 sample salaries, along with the respective means for the three treatments.

Let's consider individually the notations enclosed in parentheses in Table 12.1. There are four rows and three columns in Table 12.1. Salary observations in this table are represented by x_{ij}, where i stands for the number of rows (students) and j the number of columns (treatments). There are a total of $n \times m$ observations in the table; in this case, $4 \times 3 = 12$. For example, x_{32} denotes the salary of the third student who has 2 years of work experience. In this problem, different years of work experience are indicated in the columns of the table, and interest centers on the differences among salaries in the three columns. This is typical of *one-factor* (or *one-way*) analysis of variance, in which an attempt is made to assess the effect of only one factor (in this case, years of work experience) on the observations. Here we denote the values in the columns as x_{i1}, x_{i2}, and x_{i3} and the totals of these columns as $\Sigma_i x_{i1}$, $\Sigma_i x_{i2}$, and $\sum_i x_{i3}$. The subscript i under the summation signs indicates that the total of each column is obtained by summing the entries over the rows. We will refer to the means of the columns as $\bar{x}_1, \bar{x}_2$, and $\bar{x}_3$, or, in general, as $\bar{x}_j$. Finally, we denote the overall mean as $\bar{x}$, where $\bar{x}$ is the mean of all observations.

12.2.2 Specifying the Hypotheses

As stated earlier, we want to test whether the combination of a Bachelor's degree with 3 different levels of work experience affects beginning salaries. From Table 12.1, we calculated the following mean salaries of graduates from the three combinations: $\bar{x}_1 = \$17$, $\bar{x}_2 = \$20$, and $\bar{x}_3 = \$23$. Also included in the table is an overall average of the 12 graduates, $\bar{x} = \$20$, which is referred to as the *overall mean*. Hence, we want to test whether these three sample means were drawn from populations that have identical means. In other words, we want to test the following null hypothesis:

Table 12.1 Salaries of 12 graduates with varying work experience (in thousands of dollars)

	Years of work experience (j)		
Student (i)	*1*(x_{i1})	*2*(x_{i2})	*3* (x_{i3})
1	16 (x_{11})	19 (x_{12})	24 (x_{13})
2	2l (x_{21})	20 (x_{22})	2l (x_{23})
3	I8 (x_{3l})	21 (x_{32})	22 (x_{33})
4	I3 (x_{4l})	20 (x_{42})	25 (x_{43})
Total $\left(\sum_i x_{ij}\right)$	68$\left(\sum_i x_{i1}\right)$	80$\left(\sum_i x_{i2}\right)$	92$\left(\sum_i x_{i3}\right)$
Mean ($\bar{x}_j$)	17 ($\bar{x}_1$)	20 ($\bar{x}_2$)	23 ($\bar{x}_3$)

$$\text{Overall mean } \bar{x} = \sum_{j=1}^{3} \bar{x}_j/3$$
$$= (\$17 + \$20 + \$23)/3 = \$20$$

$$H_0 : \mu_1 = \mu_2 = \mu_3 \tag{12.1}$$

against the alternative hypothesis

H_1: At least two population means are not equal.

Thus, we are testing whether the differences between the sample means are too large to be attributed solely to chance. If the test results indicate that the sample means are significantly different, then we can conclude that different years of work experience have an impact on beginning salaries. Note that we make inferences concerning the means of more than two populations here.

12.2.3 Generalizing the One-Way ANOVA

Table 12.1 can be generalized to resemble Table 12.2, where we see n observations and m populations. Here, each of the m populations is a treatment. Table 12.2 illustrates how we initially set up a generalized matrix to perform the one-way analysis of variance. Here, the top row indicates that we will be testing the equality of m different means. Within each column, there are n individual samples taken from each of the m treatments. In developing the one-way ANOVA model, our purpose is to specify the underlying relationships among the various treatments. Hence, the first step is to calculate the sample means from the random observations taken from each of the m treatments. That is,

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}, \quad j = 1, \ldots, m \tag{12.2}$$

Table 12.2 General notation corresponding to Table 12.1

	Population (j)				
Observation (i)	*1*	*2*	*3*	...	*m*
1	x_{11}	x_{12}	x_{13}	...	x_{1m}
2	x_{21}				
3					
4					
⋮					
n	x_{n1}	x_{n2}	x_{n3}	...	x_{nm}
Total	$\sum_i x_{i1}$	$\sum_i x_{i2}$	$\sum_i x_{i3}$		$\sum_i x_{im}$
Mean	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$		$\bar{x}_m$

where

$\bar{x}_j$ — sample mean for the jth treatment
x_{ij} — ith sample observation for the jth treatment
n_j — number of sample observations in the jth treatment

As we have noted, the null hypothesis specifies that all of the j treatments have identical means. Thus, the next step in our analysis is to obtain an estimate of a common mean, which we will call the *global* mean. The *global mean*, or *overall mean*, is the summation of the individual sample observations divided by the number of total observations. Stated formally,

$$\bar{x} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} x_{ij}}{n} \tag{12.3}$$

where
$\bar{x}$ = global mean

$$n = \sum_{j=1}^{m} n_j$$

and variables x_{ij} and n_j have the same meaning as in Eq. 12.2. Alternatively, we can restate the overall mean as

$$\bar{x} = \frac{\sum_{j=1}^{m} n_j \bar{x}_j}{n} \tag{12.4}$$

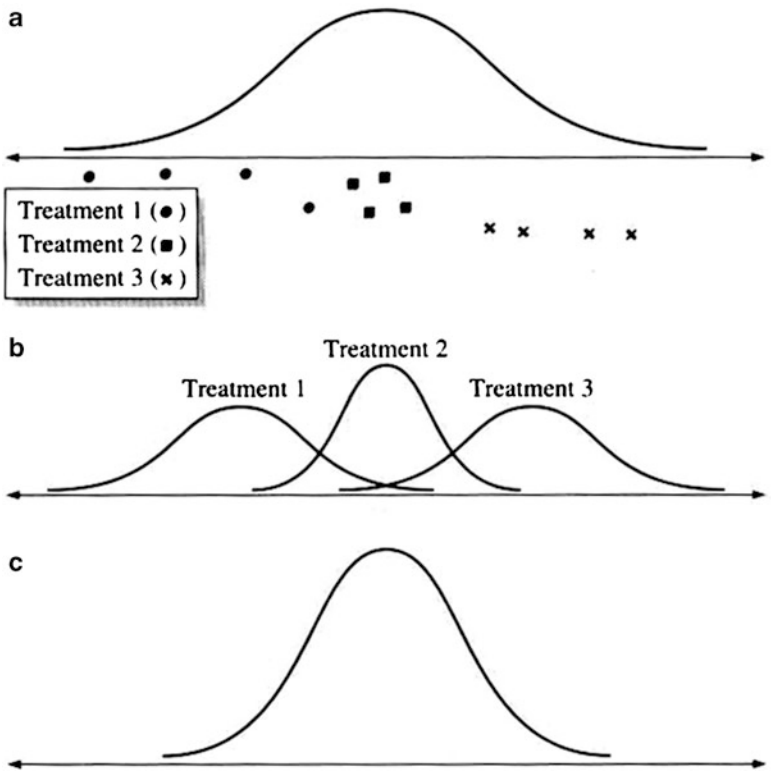


Fig. 12.1 Distributions for ANOVA: (**a**) Sample data combined, (**b**) each treatment viewed separately, (**c**) the assumption the null hypothesis makes

To test the null hypothesis that the treatment means are equal, we need to assess two measures of variability. First, we are interested in the variability of the sample within each treatment. This classification of variability is called *within-group variability*. We are also interested in the variability between the m treatments, which is called *between-groups variability*. From our sample data, we can obtain measures of both.

12.2.4 Between-Treatments and Within-Treatment Sums of Squares

The aforementioned concepts are illustrated graphically in Fig. 12.1. When sample data are combined, they appear to be observations from a single population with high dispersion, as shown in part (a). But when each treatment is viewed separately, these salary figures appear to belong to three separate populations with a smaller

variance, as indicated in part (b). Under the null hypothesis, the treatment populations have identical frequencies, as shown in part (c).

The term *variation* refers to the sum of squared deviations, which is also called the *sum of squares*. We begin our analysis of variance by measuring the variation between the treatment means. The calculation is

$$\text{SST} = \sum_{j}^{m} n_j \left(\bar{x}_j - \bar{x}\right)^2 \tag{12.5}$$

where

SST = between-treatments sum of squares (between-groups variability)
n_j = sample size of treatment j
$\bar{x}_j$ = sample mean of the jth treatment
$\bar{x}$ = overall mean

Table 12.3 illustrates calculation of the between-treatments variation for the data given in Table 12.1.

Substituting all squared-deviation values given in Table 12.3 into Eq. 12.5, we obtain the *between-treatments sum of squares* as follows:

$$\begin{aligned}\sum_{j=1}^{3} n_j \left(\bar{x}_j - \bar{x}\right)^2 &= 4(9) + 4(0) + 4(9) \\ &= 72\end{aligned}$$

Again, this measure of variability may specify why treatment means are different.

On the other hand, the within-treatment variability specifies the treatment effect. That is, the *within-treatment sum of squares* indicates the unexplained variability that is due to the random sampling process. The calculation is

$$\text{SSW} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)^2 \tag{12.6}$$

where

SSW = within-treatment sum of squares (within-group variability)
$\bar{x}_{ij}$ = value of the observation in the ith row and the jth column
$\bar{x}_j$ = mean of the jth treatment

Calculation of the within-treatment sum of squares is illustrated in Table 12.4.

Table 12.3 Worksheet for calculating between-treatments sum of squares

$n_1(\bar{x}_1-\bar{x})^2 = 4(17–20)^2 = 36$
$n_2(\bar{x}_2-\bar{x})^2 = 4(20–20)^2 = 0$
$n_3(\bar{x}_3-\bar{x})^2 = 4(23–20)^2 = 36$

Table 12.4 Worksheet for calculating within-treatment sum of squares

Treatment 1	Treatment 2	Treatment 3
$(16–17)^2 = 1$	$(19–20)^2 = 1$	$(24–23)^2 = 1$
$(21–17)^2 = 16$	$(20–20)^2 = 0$	$(21–23)^2 = 4$
$(18–17)^2 = 1$	$(21–20)^2 = 1$	$(22–22)^2 = 1$
$(13–17)^2 = \frac{16}{34}$	$(20–20)^2 = \frac{0}{2}$	$(25–23)^2 = \frac{4}{10}$

The between-treatments variation and within-treatment variation together represent the total variation of the ANOVA model. We calculate the total variation by summing the squared deviations of individual observations about the global mean. Formally, the total sum of squares can be written as

$$\text{TSS} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}\right)^2 \tag{12.7}$$

where

TSS = total sum of squares
x_{ij} = value of the observation in the ith row and the jth column
$\bar{x}$ = overall mean

We obtain the within-treatment sum of squares by substituting all squared-deviation values given in Table 12.4 into Eq. 12.6.

$$\begin{aligned}\text{SSW} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)^2 &= 34 + 2 + 10 \\ &= 46\end{aligned}$$

To put it more simply, we find the total sum of squares by adding the between-treatments variation to the within-treatment variation.

$$\begin{aligned}\text{TSS} &= \text{SST} + \text{SSW} \\ &= 72 + 46 \\ &= 118\end{aligned}$$

Even though the test of the null hypothesis for the one-way analysis of variance involves only between-treatments and within-treatment variation, it is useful to understand the relationship between total variation and its components.

12.2.5 Between-Treatments and Within-Treatment Mean Squares

The number of degrees of freedom associated with the between-treatments variation is *(m*–1). That is, because there are *m* treatments, or *m* sample means, there are *m* sums of squares used to measure the variation of these sample means around the overall mean. The overall mean is the only estimate of the population mean, so 1 degree of freedom is lost. Thus, in our example, which consists of three treatments, there are $3-1 = 2$ degree of freedom associated with the between-treatments mean square.

The number of degrees of freedom associated with the within-treatment variation is (*n*–*m*). Because there are $n\left(n = \sum_{j=1}^{m} n_j\right)$ observations, there are *m* sums of squares used to measure the within-treatment variation, with each deviation taken around its respective treatment mean. There are *m* treatment means, each an estimate of its respective population, so there is a loss of *m* degrees of freedom. Hence, in our example, which contains 12 observations and three treatment means, there are $12-3 = 9$ degree of freedom associated with the within-treatment mean square.

Again, the test of the null hypothesis is based on the assumption that all the *m* treatments have a common variance. If the null hypothesis is in fact true, then SST and SSW can be used as a basis for an estimate of a common variance. To calculate these estimates, we can now divide each of the variability measures by its number of degrees of freedom. Hence, the unbiased estimate of the *between-treatments mean square* can be obtained by dividing SST by *(m*–1) degrees of freedom:

$$\text{MST} = \text{SST}/(m-1)$$

where

$$\text{MST} = \text{between-treatments mean square (variance)}$$

In our example, the between-treatments mean square is MST = 72/2 = 36. Similarly, an unbiased estimate of the *within-treatment mean square* is found by dividing SSW by (*n*–*m*) degrees of freedom:

$$\text{MSW} = \text{SSW}/(n-m)$$

where

$$\text{MSW} = \text{within-treatment mean square (variance)}$$

In our example, the within-treatment mean square is MSW = 46/9 = 5.111. We test the null hypothesis that the population treatment means are equal by comparing the between-treatments mean square with the within-treatment mean square.

12.2.6 The Test Statistic

Comparison of the between-treatments mean square with the within-treatment mean square is performed by computing a ratio:

$$F = (\text{MST}/\text{MSW}) \tag{12.8}$$

If the null hypothesis that the population treatment means are equal were true, the ratio given in Eq. 12.8 would tend to equal 1. Alternatively, if the null hypothesis were not true, the ratio would be greater than 1 (MST generally cannot be smaller than MSW), which implies that the treatment means do differ because the between-treatments variance exceeds the within-treatment variance. In the context of our example, this would imply that different amounts of work experience do have an impact on starting salaries for graduates. The ratio for our example can be calculated as $F = 36/5.111 = 7.04$.

From this calculation, it appears that we can reject the null hypothesis that the population treatment means are equal. But first we need to determine how large the ratio must be in order for us to reject the null hypothesis. To do this, we must refer to the probability distribution of the F-distributed random variable discussed in Chap. 9 (the *F distribution*) and to the F table given as Table A6 in Appendix A at the end of the book. For our purposes, we will test the null hypothesis that the population treatment means are equal at the .05 level of significance. We refer to the F random variable as $F_{v1,v2,\alpha}$, where $v_1 = (m–1)$ is the between-treatments degrees of freedom, $v_2 = (n - m)$ is the within-treatment degrees of freedom, and α is the level of significance. When the null hypothesis is true, the F variable in Eq. 12.8 is distributed as $F_{v1,v2}$. From Table A6, we find that the critical value at the 5 % level of significance is

$$F_{2,9,.05} = 4.26$$

Thus, if the F ratio calculated for our example is greater than the critical value, then we can reject the null hypothesis that the population treatment means are equal. On the other hand, if the F ratio we calculate is less than the critical value, then we must accept the null hypothesis that the population treatment means are equal. Our sample F ratio, 7.04, is greater than the critical value, 4.26, so the null hypothesis is rejected at the .05 level of significance. We can conclude that the treatment means are significantly different. That is, work experience does affect starting salaries for graduates. A summary of this analysis of variance appears in Table 12.5, and Fig. 12.2 presents MINITAB output related to Table 12.5.

Table 12.5 Summary of one-way ANOVA table

(1) Source of variation	(2) Sum of squares	(3) Degrees of freedom	(4) Mean square
Between-treatments	72 (SST)	2 ($m - 1$)	36 (MST)
Within-treatment	46 (SSW)	9 ($n - m$)	5.111 (MSW)
$F_{2,9} = \frac{36}{5.111} = 7.04$			
$F_{2,9,.05} = 4.26$			

```
MTB > READ C1-C3
DATA> 16 19 24
DATA> 21 20 21
DATA> 18 21 22
DATA> 13 20 25
DATA> END
      4 rows read.
MTB > AOVONEWAY C1-C3

One-Way Analysis of Variance

Analysis of Variance
Source     DF        SS        MS        F        P
Factor      2     72.00     36.00     7.04    0.014
Error       9     46.00      5.11
Total      11    118.00
                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level       N      Mean     StDev  ---------+---------+---------+-------
C1          4    17.000     3.367  (-------*------)
C2          4    20.000     0.816           (------*------)
C3          4    23.000     1.826                   (-------*------)
                                  ---------+---------+---------+-------
Pooled StDev =     2.261           17.5      21.0      24.5
```

Fig. 12.2 MINITAB output for Table 12.5

12.2.7 Population Model for One-Way ANOVA

The one-factor ANOVA model discussed in this section can also be described in a different type of specification. Let the random variable X_{ij} denote the ith observation from the jth population, and let μ_j denote the mean of this population. In addition, let μ denote the overall mean of m combined populations.

Then the population model for ANOVA states that any value x_{ij} is the sum of the grand mean μ, the *treatment effect* τ_j, and the random error. In symbols, the one-factor ANOVA model is

$$\begin{aligned} X_{ij} &= \mu + \tau_j + \varepsilon_{ij} \\ &= \mu + (\mu_j - \mu) + (X_{ij} - \mu_j) \end{aligned} \tag{12.9}$$

where

x_{ij} = value of the dependent variable in the ith row and the jth column; this is the variable under investigation
μ_j = mean of the jth column; this is the average value for the jth treatment. $\mu_j = \mu + \tau_j$
μ = grand mean; this is the mean of all the column means
τ_j = treatment effect for the jth column, defined as $(\mu_j - \mu)$; this is the difference between a column mean and the grand mean. The value of τ_j indicates how much effect a particular treatment has on the grand mean, $\tau_j = \mu_j - \mu$
e_{ij} = random error associated with X_{ij}, defined as the difference between X_{ij} and μ_j. This is the amount by which a particular value of the dependent variable differs from the mean of all values in that column. $e_{ij} = X_{ij} - \mu_j$

By using the model of Eq. 12.9, we can redefine the null hypothesis defined in Eq. 12.1 as Eq. 12.1. Then our null hypothesis is that every population mean μ_j is the same as the overall mean μ.

$$H_0 : \tau_1 = \tau_2 = \tau_3 = 0 \tag{12.10}$$

where $\tau_j = \mu_j - \mu$ $(j = 1, 2, 3)$.

12.3 Simple and Simultaneous Confidence Intervals

In the salary study we have examined in this chapter, the analysis of variance was used to determine whether there was a difference in average salary among workers with different numbers of years of work experience. Once differences in the means of the groups are found, however, it is important to determine which particular groups are different. In other words, we are interested in establishing a confidence interval for the difference between two population means.

12.3.1 Simple Comparison

To compare the differences of the population means of group 1 and group 2, we can construct a confidence interval for $(\mu_1 - \mu_2)$ by using our estimates $(\overline{X}_1 - \overline{X}_2)$ as discussed in Chaps. 10 and 11. Formally, the $(1-\alpha)$ percent confidence interval for $\mu_1 - \mu_2$ is

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{\alpha/2,(n-m)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{12.11}$$

where

μ_1, μ_2 = population means for treatments 1 and *2,* respectively
$\bar{X}_1, \bar{X}_2$ = sample means for treatments 1 and 2, respectively
$t_{\alpha/2,(n-m)} = t$ statistic at the $\alpha/2$ level of significance with $(n–m)$ degrees of freedom
S_p^2 = SSW/*(n–m),* the within-treatment mean square, where SSW has been defined in Eq. 12.6
n_1, n_2 = number of observations for treatments 1 and 2, respectively.

Hence, our pooled variance from the three treatments is calculated as follows:

$$S_p^2 = \frac{1}{(n-m)}(\text{SSW}) = \frac{1}{9}(34 + 2 + 10)$$
$$= 5.111$$

and the pooled standard deviation is $S_p = \sqrt{S_p^2} = \sqrt{5.111} = 2.261$. As we noted earlier for the within-treatment variation, the pooled standard deviation has (*n*–m), or 9, degrees of freedom. From Table A4 in Appendix A, we have $t_{.025,9} = 2.262$. Therefore, according to Eq. 12.11, a 95 % confidence interval for the difference of the population means for treatments 1 and 2 can be determined as follows:

$$(17 - 20) \pm (2.262)(2.261).\left(\sqrt{\frac{1}{4} + \frac{1}{4}}\right) = -3 \pm 3.616 \text{ or } (-6.616, +.616)$$

Thus, we conclude that the mean salary for treatment 2 is approximately between $6,616 higher and $616 less than that for treatment 1.

Accordingly, for our example, the 95 % confidence intervals for the difference between two population treatment means are

$$(17 - 20) - 3.616 < \mu_1 - \mu_2 < (17 - 20) + 3.616, \text{ or } (-6.616, +.616)$$
$$(17 - 23) - 3.616 < \mu_1 - \mu_3 < (17 - 23) + 3.616, \text{ or } (-9.616, -2.384)$$
$$(20 - 23) - 3.616 < \mu_2 - \mu_3 < (20 - 23) + 3.616, \text{ or } (-6.616, +.616)$$

Note that each confidence interval has the same width. This is due to the fact that each interval contains the pooled variance and each treatment contains the same number of observations. Also note that not all 3 intervals overlap zero.

There is one problem with this approach. Although we may be 95 % confident of the individual intervals listed, we are less confident that the whole *system* of intervals

is true. The problem we face is to determine a simultaneous confidence level for the whole system given that the intervals are independent (all have the same S_p). We can achieve this goal by using Scheffé's multiple comparison.

12.3.2 Scheffé's Multiple Comparison

The problem we have just posed can be restated as determining how much wider the intervals must become in order for each interval simultaneously to yield a (1–α) percent level of confidence. This can be done by employing *Scheffé's multiple comparison*. For the 95 % confidence interval for the difference between the population means for treatment 1 and treatment 2, Scheffé's multiple comparison formula[1] is

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm \sqrt{(m-1)\left(F_{\alpha,m-1,n-m}\right)\left(S_p^2\right)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \qquad (12.12)$$

where

$F_{\alpha,m-1,n-m}$ = critical value of F with (m–1) and (n–m) degrees of freedom at the α level of significance
m = number of the means to be compared
n_1, n_2 = number of observations for combination 1 and combination 2, respectively
S_p = sample standard deviation pooled from all samples

From Table A6, we have $F_{.05,2,9} = 4.26$. Then, at a 95 % level of confidence,

$$(17-20) \pm \sqrt{2(4.26)(2.261)^2}\sqrt{\left(\frac{1}{4}\right)+\left(\frac{1}{4}\right)} = -3 \pm (2.919)(2.261)(.707)$$
$$= -3 \pm 4.666, \text{ or } (-7.666, +1.666)$$

For the entire system, we have

$$-7.666 < \mu_1 - \mu_2 < +1.666$$
$$-10.666 < \mu_1 - \mu_3 < -1.334$$
$$-7.666 < \mu_2 - \mu_3 < +1.666$$

[1] See H. Scheffé(l959), *The Analysis of Variance* (New York: Wiley). This method has been adjusted for the number of means to be compared. This is the simplest case of Scheffé's multiple comparison.

As we expected, the increased width of each interval now makes us 95 % confident that all the foregoing statements are simultaneously true. Again, not all 3 intervals overlap zero; only the second null hypothesis should be rejected.

12.4 Two-Way ANOVA with One Observation in Each Cell, Randomized Blocks

In this section, we extend one-way ANOVA to two-way ANOVA. We discuss first the case of two-way ANOVA with one observation per cell, then the case of two-way ANOVA with more than one observation per cell.

12.4.1 Basic Concept

This section offers a more in-depth interpretation of ANOVA technique. In the example we have been using, our primary interest focused on a single aspect of the one-way analysis of variance (years of work experience), but it is possible that another factor also affects the outcome. In the one-way analysis of variance, we concluded that number of years of work experience had a significant impact on starting salary. However, we may suspect that some of the variability of the model is due to the geographic location of the job. Hence, we now want not only to look at *treatment effects* of number of years of work experience but also to isolate the impact of geographic location on the starting salaries of the graduates. By setting up a two-way ANOVA problem, we want to design a more accurate test to explain the differences in the mean population of the various treatments.

Our new model must be constructed in such a way as to test for the influence that a second factor may have on the starting salaries. Using the data from Table 12.6, we will have the 4 rows represent 4 geographic locations in the United States. Hence, we will be able to acquire information about the various years of work experience as well as information about the geographic location of the job. This new factor in our analysis is called a *blocking variable*. To simplify our analysis, the blocks will contain only a single observation per cell. Thus, as in our one-way ANOVA problem, we will use only the 12 observations from Table 12.6. Each of the four rows will represent a geographic location.

Row 1: West	Row 3: Northeast
Row 2: Midwest	Row 4: South

Table 12.6 illustrates how to set up the two-way analysis of variance. The salary data of Table 12.6 are identical to those of Table 12.1. However, in Table 12.6, we interpret 4 students' salaries within each column, representing salaries from 4

Table 12.6 Salaries of 12 students with varying work experience in 4 different geographic locations (in thousands of dollars)

	Years of work Experience				
Region	*1*	*2*	*3*	Row sums	Row means
1	16	19	24	59	19.667
2	21	20	21	62	20.667
3	18	21	22	61	20.333
4	13	20	25	58	19.333
Column sums	68	80	92	240	
Column means	17	20	23		
		Global mean $\bar{x} = 20$			

different geographic regions. Therefore, different locations (regions) constitute an additional factor.

12.4.2 Specifying the Hypotheses

Our purpose is to test the following two hypotheses:

1. H_0: Population mean salaries among various years of work experience are equal.
2. H_0: Population mean salaries among various geographic locations are equal.

Again, the alternative hypotheses are that the mean population values are not equal.

12.4.3 Between and Residual Sum of Squares

The necessary calculations for the two-way analysis of variance are:

SST = between-treatments sum of squares
SSB = between-blocks sum of squares
TSS = total sum of squares
SSE = error sum of squares

Here treatments and blocks represent different years of work experience and different locations, respectively.[2]

From the one-way analysis of variance, we have already calculated SST (Table 12.3), SSW (Table 12.4), and TSS. The next step is to calculate the

[2] Alternatively, we can use levels of factors (treatments) A and B to represent different years of work and different locations.

between-blocks (between-rows) sum of squares. In this case, the observation can be represented by x_{ijk}. The subscripts i, j, and k represent the kth salary observation in the ith row and the jth column. Then the between-rows sum of squares can be defined as

$$\text{SSB} = \sum_{i=1}^{I} JK(\bar{x}_{i.} - \bar{x})^2 \tag{12.13}$$

where

SSB = between-blocks sum of squares

$$\bar{x}_{i..} = \text{sample mean of the } i\text{th row} = \frac{\sum_{j=1}^{J} x_{ij.}}{JK}$$

$\bar{x}$ = overall mean

Table 12.7 illustrates calculation of the between-blocks sum of squares for our example.

The between-treatments (between columns) sum of squares can be defined as

$$\text{SST} = \sum_{j=1}^{J} IK\left(\bar{x}_{.j.} - \bar{x}\right)^2 \tag{12.14}$$

where

SST = between-treatments sum of squares

$$\bar{x}_{.j.} = \text{sample mean of the } j\text{th column} = \frac{\sum_{i=1}^{I} x_{ij.}}{IK}$$

$\bar{x}$ = over all mean

From Table 12.3, we have SST = 72.

Finally, because TSS = SST + SSB + SSE, the residual sum of squares is calculated as follows:

$$\text{SSE} = \text{TSS} - \text{SST} - \text{SSB} = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ijk} - \bar{x}_i - \bar{x}_j + \bar{x}\right)^2 \tag{12.15}$$

Table 12.7 Between-blocks sum of squares

$(J)(\bar{x}_1-\bar{x}) = (3)(19.667-20)^2 = .333$
$(J)(\bar{x}_2-\bar{x}) = (3)(20.667-20)^2 = 1.335$
$(J)(\bar{x}_3-\bar{x}) = (3)(20.333-20)^2 = .333$
$(J)(\bar{x}_4-\bar{x}) = (3)(19.333-20)^2 = 1.335$
SSB = 3.336

Hence, SSE = 118−72−3.336 = 42.664.

Before we can proceed with the test of our hypotheses, we must determine how many degrees of freedom are associated with the between-blocks variation and the residual variation.

The number of degrees of freedom associated with the between-blocks variation is (*I*–1). That is, because there are *I* blocks, or *I* sample factor means, there are *n* sums of squares used to measure the variation of these sample means around the global mean. The global mean is again the only estimate of the population mean, so 1 degree of freedom is lost. Thus, for our example, which consists of 4 levels of the block, there are 4−1 = 3 degree of freedom associated with the between-blocks sum of squares.

12.4.4 Between Variance, Error Variance, and F-Test

The number of degrees of freedom associated with the residual variation is (*J*–1) (*I*–1). In this instance, the residual variation takes into account both the variation between the treatments and the variation between the blocks. Hence, we must adjust the residual variation by the degrees of freedom associated with both the between-treatments degrees of freedom and the between-blocks degrees of freedom. Thus, for our example, the number of degrees of freedom associated with the residual variation is (2)(3) = 6.

Now we can obtain unbiased estimates of the between-blocks variance and the residual variance. The between-blocks variance is calculated as follows:

$$\text{MSB} = \frac{\text{SSB}}{(I-1)} = \frac{3.336}{3} = 1.112$$

Analogously, the residual variance is

$$\text{MSE} = \frac{\text{SSE}}{(J-1)(I-1)} = \frac{42.664}{6} = 7.111$$

To test our null hypothesis about the influence of various years of work experience, we must calculate the *F* ratio.

$$F(2,6) = \frac{\text{MST}}{\text{MSE}} = \frac{36}{7.111} = 5.063$$

The critical value associated with this test is 5.14 ($F_{2,6,.05}$), from Table A6 in Appendix A at the end of the book. Because the F ratio is less than the critical value, we cannot reject the null hypothesis that there is no difference between the population means of salaries associated with various years of work experience.

In testing the null hypothesis for the influence of geographic location on salaries, we find that the F ratio is

$$F(3,6) = \frac{\text{MSB}}{\text{MSE}} = \frac{1.112}{7.111} = .156$$

The critical value associated with this test is 4.76 ($F_{3,6,.05}$), from Table A5. Again we cannot reject the null hypothesis that the population means of salaries associated with geographic location are equal.

In conclusion, having accepted both hypotheses, we can state that there are no significant differences among various years of work experience or among various geographic locations in the effect they have on starting salaries for graduates. Note that the effect of work experience obtained from two-way ANOVA is different from that of one-way discussed in Sect. 12.2. Table 12.8 summarizes the data for the two-way analysis of variance. The MINITAB output of Table 12.8 is presented in Fig. 12.3.

Because $F_{2,6} = 5.06 < 5.14$, we conclude that the null hypothesis cannot be rejected. In other words, different years of work experience do not affect starting salary. Similarly, $F_{3,6} = .156 < 4.76$, so we should conclude that no salary differences exist among different regions. This is a good illustration of the fact that MSB is generally smaller than MSE when the null hypothesis is true.

For the two-factor model, we use three subscripts. As in the one-factor model, the letter j represents column treatments and runs from 1 to J. The letter i represents row treatments and runs from 1 to I. The letter k represents the number of the observations in a cell and runs from 1 to K.

12.4.5 *Population Model for Two-Way ANOVA with One Observation in Each Cell*

Following Eq. 12.9 and assuming there is no interaction between treatment and block, we can construct a population model for two-way ANOVA without interaction. It is

$$X_{ijk} = \mu + \tau_j + \lambda_i + \varepsilon_{ijk} \tag{12.16}$$

where

$X_{ijk} = k$th population value in the jth column and the ith row
$\mu_j =$ population mean of the jth treatment

Table 12.8 Two-way ANOVA summary

(1) Source of variation	(2) Sum of squares	(3) Degrees of freedom	(4) Mean square
Between-treatments	72 (SST)	2 ($J-1$)	36
Between-blocks	3.336 (SSB)	3 ($I-1$)	1.112
Residuals	42.664 (SSE)	6 [$(J-1)(I-1)$]	7.111

$\text{MST} = \frac{\text{SST}}{J-1} = \frac{72}{2} = 36$
$\text{MSB} = \frac{\text{SSB}}{I-1} = \frac{3.336}{3} = 1.112$
$\text{MSE} = \frac{\text{SSE}}{(J-1)(I-1)} = \frac{42.664}{6} = 7.111$
$F_{2,6} = \frac{\text{MST}}{\text{MSE}} = \frac{36}{7.111} = 5.063$
$F_{2,6,.05} = 5.14$
$F_{3,6} = \frac{\text{MSB}}{\text{MSE}} = \frac{1.112}{7.111} = .156$
$F_{3,6,.05} = 4.76$

```
MTB > READ C1-C3
DATA> 16 1 1
DATA> 21 1 2
DATA> 18 1 3
DATA> 13 1 4
DATA> 19 2 1
DATA> 20 2 2
DATA> 21 2 3
DATA> 20 2 4
DATA> 24 3 1
DATA> 21 3 2
DATA> 22 3 3
DATA> 25 3 4
DATA> END
     12 rows read.
MTB > TWOWAY USING DATA IN C1, LEVEL IN C2, BLOCK IN C3

Two-way Analysis of Variance


Analysis of Variance for C1
Source        DF         SS         MS
C2             2      72.00      36.00
C3             3       3.33       1.11
Error          6      42.67       7.11
Total         11     118.00
```

Fig. 12.3 MINITAB output for Table 12.8

μ_i = population mean of the ith block
μ = grand mean of the population
λ_i = block effect of the ith row; $\lambda_i = \mu_i - \mu$
τ_j = block effect of the jth column; $\tau_j = \mu_j - \mu$

So far we have discussed two-way ANOVA with only one observation in each cell. If there is more than one observation in each cell, then there exists an interaction effect in addition to treatment and block effects. And matters get more complicated.

12.5 Two-Way ANOVA with More than One Observation in Each Cell

12.5.1 Basic Concept and Hypothesis Testing

The data that we used to do the two-way ANOVA contained only one observation in each cell. Now we expand the data set of Table 12.6 by allowing two sample observations in each cell, as shown in Table 12.9. Here the total sums of squares can be dissected into four components and can be defined as follows:

Total sums of squares (TSS) = between-treatments sum of squares (SST) + between-blocks sum of squares (SSB) + interaction sum of squares (SSI) + error sum of squares (SSE)

or, more briefly,

$$\text{TSS} = \text{SST} + \text{SSB} + \text{SSI} + \text{SSE}$$

In this case, we add an interaction sum of squares because there is more than one observation in each cell. On the basis of the data listed in Table 12.9, we calculate block means, treatment means, cell means, and overall mean as follows:

1. Block means

$$\bar{x}_{1..} = \frac{16 + 16.5 + \ldots + 25}{6} = 19.583$$
$$\bar{x}_{2..} = \frac{21 + 20.5 + \ldots + 22.5}{6} = 20.667$$
$$\bar{x}_{3..} = \frac{18 + 19 + \ldots + 21}{6} = 20.317$$
$$\bar{x}_{4..} = \frac{13 + 13.5 + \ldots + 23}{6} = 19.217$$

Table 12.9 Salaries of 24 students with varying work experience in four different geographic locations (in thousands of dollars)

	Years of work experience					
Region	*1*		*2*		*3*	
1	16	16.5	19	17	24	25
2	21	20.5	20	19	21	22.5
3	18	19	21	20.9	22	21
4	13	13.5	20	20.8	25	23

2. Treatment means

$$\bar{x}_{.1.} = \frac{16 + 21 + \ldots + 13.5}{8} = 17.188$$
$$\bar{x}_{.2.} = \frac{19 + 20 + \ldots + 20.8}{8} = 19.713$$
$$\bar{x}_{.3.} = \frac{24 + 21 + \ldots + 23}{8} = 22.938$$

3. Cell means

$$\bar{x}_{11.} = \frac{16 + 16.5}{2} = 16.25 \qquad \bar{x}_{21.} = \frac{19 + 17}{2} = 18$$

Similarly, we can obtain

$$\bar{x}_{21.} = 20.75 \quad \bar{x}_{22.} = 19.50 \quad \bar{x}_{13.} = 24.50 \quad \bar{x}_{43.} = 24.00$$
$$\bar{x}_{31.} = 18.50 \quad \bar{x}_{32.} = 20.95 \quad \bar{x}_{23.} = 21.75$$
$$\bar{x}_{41.} = 13.25 \quad \bar{x}_{42.} = 20.40 \quad \bar{x}_{33.} = 21.50$$

4. Overall mean

We use the average of column means to calculate the overall mean.

$$\bar{x} = \frac{17.188 + 19.713 + 22.938}{3} = 19.946$$

Using all related data, we calculate TSS, SST, and SSB as follows:

$$\begin{aligned}
\text{TSS} &= \sum_i \sum_j \sum_k \left(x_{ijk} - \bar{x}\right)^2 = (16 - 19.946)^2 \\
&\quad + (21 - 19.946)^2 + \cdots + (23 - 19.946)^2 \\
&= 226.380 \\
\text{SST} &= IK \sum_{j=1}^{J} (x_{ijk} - \bar{x})^2 = (4)(2)\Big[(17.188 - 19.946)^2 \\
&\quad + (19.713 - 19.946)^2 + (22.938 - 19.946)^2\Big] \\
&= 132.903 \\
\text{SSB} &= JK \sum_{i=1}^{I} (\bar{x}_{i..} - \bar{x})^2 = (3)(2)\Big[(19.583 - 19.946)^2 \\
&\quad + (20.667 - 19.946)^2 + (20.317 - 19.946)^2 \\
&\quad + (19.217 - 19.946)^2\Big] \\
&= 7.921
\end{aligned}$$

Because there is more than one observation in each cell, the SSE given by Eq. 12.15 can be dissected into interaction and error. The interaction term (SSI) is similar to the SSE with only one observation in each cell, as defined in Eq. 12.15. The error term, SSE, can be defined as

$$\text{SSE} = \sum_i \sum_j \sum_k \left(x_{ijk} - \bar{x}_{ij.}\right)^2 \tag{12.17}$$

In terms of our data, the SSI and SSE are calculated as follows:

$$\begin{aligned}
\text{SSI} &= K \sum_i \sum_j \left(\bar{x}_{ij.} - \bar{x}_{.j.} - \bar{x}_{i..} + \bar{x}\right)^2 \\
&= 2\Big[(16.250 - 17.188 - 19.583 + 19.946)^2 \\
&\quad + (20.750 - 17.188 - 20.667 + 19.946)^2 \\
&\quad + \cdots + (21 - 22.938 - 20.317 \\
&\quad + 19.946)^2 + (24 - 22.938 - 19.217 + 19.946)^2\Big] \\
&= 77.730 \\
\text{SSE} &= \sum_i \sum_j \sum_k \left(x_{ijk} - \bar{x}_{ij.}\right)^2 \\
&= (16 - 16.250)^2 + (16.500 - 16.250)^2 \\
&\quad + \ldots + (25 - 24)^2 + (23 - 24)^2 \\
&= 7.825
\end{aligned}$$

Table 12.10 Two-way analysis of variance with interaction summary

(1) Source of variation	(2) Sum of squares	(3) Degrees of freedom	(4) Mean square	(5) F ratio
Between-treatments (SST)	132.903	2	66.452	101.91
Between-blocks (SSB)	7.921	3	2.640	4.05
Interaction (SSI)	77.730	6	12.955	19.87
Errors (SSE)	7.825	12	.652	
Total (TSS)	226.379			

Using the foregoing data, we calculate the two-way ANOVA table with interaction. Our results are listed in Table 12.10.

From the F ratio shown in column (5) of Table 12.10, we can test whether the years-of-education effect, the regional effect, and the interaction effect are statistically significant. From Table A6 in Appendix A, we find that the critical values at $\alpha = .05$ are $F_{2,12,.05} = 3.89$, $F_{3,12,.05} = 3.49$, and $F_{6,12,.05} = 3.00$. Comparing these values with the F ratio listed in column (5) of Table 12.10 leads to the conclusion that number of years of education, geographic region, and their interaction all have significant impacts on the starting salary. The MINITAB output of Table 12.10 is presented in Fig. 12.4.

12.5.2 Generalizing the Two-Way ANOVA

If there are several observations per cell, then the cell mean can be defined as

$$\bar{x}_{ij.} = \frac{\sum_{k=1}^{K} x_{ijk}}{K}$$

Here the column (group) mean $\bar{x}_{.j.}$ and the row (block) mean $\bar{x}_{i..}$ can be defined as

$$\bar{x}_{.j.} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{K} x_{ijk}}{IK}$$

$$\bar{x}_{i..} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}}{JK}$$

```
MTB > READ C1-C3
DATA> 16 1 1
DATA> 16.5 1 1
DATA> 21 1 2
DATA> 20.5 1 2
DATA> 18 1 3
DATA> 19 1 3
DATA> 13 1 4
DATA> 13.5 1 4
DATA> 19 2 1
DATA> 17 2 1
DATA> 20 2 2
DATA> 19 2 2
DATA> 21 2 3
DATA> 20.9 2 3
DATA> 20 2 4
DATA> 20.8 2 4
DATA> 24 3 1
DATA> 25 3 1
DATA> 21 3 2
DATA> 22.5 3 2
DATA> 22 3 3
DATA> 21 3 3
DATA> 25 3 4
DATA> 23 3 4
DATA> END
     24 rows read.
MTB > TWOWAY USING DATA IN C1, A LEVEL IN C2, B LEVEL IN C3

Two-way Analysis of Variance


Analysis of Variance for C1
Source          DF         SS         MS
C2               2    132.903     66.452
C3               3      7.921      2.640
Interaction      6     77.730     12.955
Error           12      7.825      0.652
Total           23    226.380
```

Fig. 12.4 MINITAB output for Table 12.10

In addition, the overall mean ($\bar{x}$) can be defined as

$$\bar{x} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}/IJK = \sum_{j=1}^{J} \bar{x}_{.j.}/J = \sum_{i=1}^{I} x_{i..}/I$$

Using the cell mean, treatment mean, block mean, overall mean, and other related concepts and notations discussed in this section, we can define the general format of the two-way ANOVA with K observations per cell as shown in Table 12.11. The population model for Table 12.11 is

$$X_{ijk} = \mu + \tau_j + \lambda_i + (\lambda\tau)_{ij} + \varepsilon_{ijk} \tag{12.18}$$

where μ, τ_j λ_i, and ϵ_{ijk} are as defined in Eq. 12.16

Table 12.11 General format of the two-way ANOVA table with K observation per cell

(1) Source of variation	(2) Sum of squares	(3) Degrees of freedom	(4) Mean square	(5) F ratio
Between-treatments	$\text{SST} = IK\sum_{j=1}^{J}\left(\bar{x}_{.j.} - \bar{\bar{x}}\right)^2$	$J-1$	$\text{MST} = \frac{\text{SST}}{J-1}$	$\frac{\text{MST}}{\text{MSE}}$
Between-blocks	$\text{SSB} = JK\sum_{i=1}^{I}\left(\bar{x}_{i..} - \bar{\bar{x}}\right)^2$	$I-1$	$\text{MSB} = \frac{\text{SSB}}{I-1}$	$\frac{\text{MSB}}{\text{MSE}}$
Interaction	$\text{SSI} = K\sum_i\sum_j\left(\bar{x}_{ij.} - \bar{x}_{.j.} - \bar{x}_{i..} + \bar{\bar{x}}\right)^2$	$(J-1)(I-1)$	$\text{MSI} = \frac{\text{SSI}}{(J-1)(I-1)}$	$\frac{\text{MSI}}{\text{MSE}}$
Error	$\text{SSE} = \sum_i\sum_j\sum_k\left(x_{ijk} - \bar{x}_{ij.}\right)^2$	$JI(K-1)$	$\text{MSE} = \frac{\text{SSE}}{IJ(K-1)}$	
Total	$\text{TSS} = \sum_i\sum_j\sum_k\left(x_{ijk} - \bar{\bar{x}}\right)^2$	$IJK-1$		

$(\lambda\tau)_{ij}$ = interaction effect in the ith row and the jth column = $\mu_{ij}-\mu_i-\mu_j + \mu$
μ_{ij} = population mean of the cell in the ith row and the jth column
μ_i = population mean of the ith block
μ_j = population mean of the jth treatment

12.6 Chi-Square as a Test of Goodness of Fit

In this section, we will show how the chi-square statistic can be used to test the appropriateness of a distribution—its goodness of fit for a set of data. *Goodness-of-fit tests* are designed to study the frequency distribution to determine whether a set of data are generated from a certain probability distribution, such as the uniform, binomial, Poisson, or normal distribution.

Among the goodness-of-fit tests, the *chi-square test* is used to test the equality of more than two proportions if a probability distribution is assumed to be uniform. This is similar to using the F statistic to test the equality of more than two means in the analysis of variance.

If a marketing manager is interested in knowing whether 4 different brands of painkillers are recommended equally often by doctors (or enjoy the same market shares), the manager can set up the following hypotheses:

H_0: Same market shares
H_1: Different market shares

To test this hypothesis, the manager can send out questionnaires to 1,000 doctors asking what painkiller they usually recommend to their patients. The responses can be tallied to obtain the observed sample frequency distribution. The tallied

responses are called the *observed frequency*. If the null hypothesis of equal market shares is true, we would expect to see that roughly 250 doctors recommended each brand. This frequency distribution is called the *expected frequency* because it is anticipated when the null hypothesis is true. In applying the goodness-of-fit test, we compare the expected frequency with the observed frequency to determine whether the observed frequency conforms to the expected frequency—and hence supports the null hypothesis. If the null hypothesis is true, the frequencies for four brands of painkillers will be equal. Therefore, we can regard this example as a test of uniform distribution.

To take another example, many statistical inferences drawn in studying stock rates of return are based on the assumption that the rates of return of a stock follow a normal distribution. It should be interesting to test whether the rates of return are really generated from a normally distributed population. Here the null hypothesis is that the data are from a normally distributed population, and the alternative hypothesis is that the data are not from a normally distributed population. Again we perform the goodness-of-fit test by comparing the anticipated frequency distribution when the null hypothesis is true with the frequency distribution that is actually observed.

The chi-square statistic for determining whether the data follow a specific probability distribution is

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{\left(f_i^o - f_i^e\right)^2}{f_i^e} \tag{12.19}$$

where

f_i^o = observed frequency
f_i^e = expected frequency
k = number of groups
χ^2_{k-1} = chi-square statistic with (k–1) degrees of freedom

The following examples illustrate various applications of the goodness-of-fit test in deciding whether the population that generates the data follows a presumed distribution. This presumed distribution can be a uniform distribution, a binomial distribution, or a Poisson distribution.

Example 12.1 Uniform Distribution: Market Shares of Different Types of Cars. A marketing manager wants to test his belief that four different categories of cars share the auto market equally. These four categories of cars are brand A, brand B, brand C, and imported cars. He sends out 2,000 questionnaires to car owners throughout the nation and receives the following responses:

Brand	Number of owners (observed frequency)
A	475
B	505
C	495
Imported	525
Total	2,000

Armed with these data, we can help him solve the problem. We first set up the hypotheses

H_0 : Same market shares(uniform distribution)

H_1 : Different market shares (nonuniform distribution)

The chi-square test statistic defined in Eq. 12.19 can be used to perform the hypothesis test. When the null hypothesis is true, there should be 500 responses for each category of product. This implies that the expected frequency should be 500 for each category of product. Computation of the chi-square statistic in terms of Eq. 12.19 is given in column (4) of Table 12.12.

In this example, we divided the total sample into four groups. The frequencies of these four groups must add up to 2,000. This means that when any three groups' frequencies are known, the fourth group's frequency is also set. The number of degrees of freedom is therefore $(k–1)$, so here it is $4-1 = 3$. From the χ^2 distribution table (Table A5 in Appendix A of this book), we obtain $\chi^2_{3,.05} = 7.81$. Because

$$\chi^2_3 = \sum_{i=1}^{k} \frac{\left(f_i^o - f_i^e\right)^2}{f_i^e} = 2.6$$

which is smaller than 7.81, we fail to reject the null hypothesis at $\alpha = .05$. We conclude that we do not have enough evidence to argue that the frequency distribution of different car brands is not uniformly distributed. In other words, the differences in market share among these four different brands of automobiles are not statistically significant.

Example 12.2 Binomial Distribution: Correct Picks in a Football Pool. A football fan keeps track of the football betting record for the football betting pool in her company. In each bet, a player has to pick the winner for ten games. In the last season, 1,000 bets were placed. The numbers of correct picks are tallied in column (2) of Table 12.13; these figures are observed frequencies.

We would like to know whether the numbers of correct picks follow a binomial distribution with $P = .5$. Accordingly, we have

H_0: A binomial distribution with $P = .5$ is a good description of the number of correct picks.

Table 12.12 Computation of the chi-square test statistic for Example 12.1

(1) Brand	(2) Number of owners	(3) f_i^e	(4) $(f_i^o - f_i^e)^2 / f_i^e$
A	475	500	5/4
B	505	500	1/20
C	495	500	1/20
Imported	525	500	5/4
Sum	2000	2000	$\chi_3^2 = 52/20 = 2.6$

Table 12.13 Computation of the chi-square test statistic for football betting pool problem

(1) Number of correct picks	(2) Number of bets, f_i^o	(3) Expected binomial probability	(4) Expected frequency, f_i^e	(5) $(f_i^o - f_i^e)^2 / f^e$
0	2	.001	1	1
1	8	.010	10	.4
2	39	.044	44	.57
3	123	.117	117	.31-
4	207	.205	205	.02
5	250	.246	246	.07
6	203	.205	205	.02
7	115	.117	117	.03
8	40	.44	44	.36
9	13	.10	10	.9
10	0	.001	1	1
Sum	1,000	1.00	1,000	4.68

H_1: A binomial distribution with $P = .5$ is not a good description of the number of correct picks.

To solve this problem, we must determine whether the discrepancies between the observed frequencies and those we would expect to observe if the binomial distribution were the proper model to use are actually due to chance. To calculate the expected frequencies, we find the probabilities of the numbers of correct picks in Table Al in Appendix A by looking for $n = 10$ and $P = .5$. The probabilities are listed in column (3) of Table 12.13. Since the number of bets is 1,000, the expected frequencies f_i^e can be obtained by multiplying the probabilities listed in column (3) by 1,000; they are indicated in column (4). Again, comparing the observed and expected frequencies gives us the test statistic. Column (5) of Table 12.13 gives the results of computation of the test statistic in accordance with Eq. 12.19. From column (5), we obtain

$$\sum_i^k \left(f_i^o - f_i^e\right)^2 / f_i^e = 4.68$$

Table 12.14 Computation of the chi-square test statistic for patient arrivals

	Number of patients arriving during 1 h				
	0	*1*	*2*	*3 or more*	*Sum*
(1) Number of hours, f_i^o	60	140	125	155	480
(2) Probability	.135	.271	.271	.323	1.00
(3) Number of hours, f_i^e	65	130	130	155	480
(4) $\frac{(f_i^o - f_i^e)^2}{f_i^e}$	.38	.77	.19	0	1.34

From Table A5, we find that $\chi^2_{10,.05} = 18.31$. Because $4.68 < 18.31$, we conclude that there is not enough evidence for us to reject, at $\alpha = .05$, the null hypothesis that the data are from a binomial distribution.

Example 12.3 Poisson Distribution: Number of Patient Arrivals. Suppose a hospital has kept track of the number of patients arriving at the emergency room during a given hour for the last 480 h (20 days). It was found that 960 patients came to the emergency room during that period. The observed distribution of the arrival of patients is given in row (1) of Table 12.14. We would like to know whether this distribution is a Poisson distribution. If the null hypothesis is true, the data are generated by the Poisson probability distribution—that is,

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where λ is the expected number of patients arriving in a given hour. In the last 480 h, there were 960 patients. The expected number of patients, λ, can be estimated as 2(960/480) per hour. Using the foregoing Poisson distribution formula, we compute the probability of $x = 0,1,2$, and $x \geq 3$. We use the Poisson probability table, Table A2 in Appendix A, to obtain the expected probabilities indicated in row (2) of Table 12.14. Multiplying the probabilities by 480, we obtain f_i^e as shown in row (3) of Table 12.14. Row (4) of this table gives the values of $\left(f_i^o - f_i^e\right)^2 / f_i^e$. From row (4), we obtain

$$\chi^2_{4-1} = .38 + .77 + .19 + 0 = 1.34 < \chi^2_{3,.05} = 7.81$$

Therefore, we conclude that the null hypothesis that patient arrivals are generated by the Poisson probability distribution cannot be rejected.

12.7 Chi-Square as a Test of Independence

In this section, we show how to use the chi-square test introduced in Sect. 12.6 to test the independence of two variables (this was briefly discussed in Chap. 5). Suppose a sample is taken from a population each of whose members can be

Table 12.15 300 students classified by grades and major

	Grade				
Major	*A*	*B*	*C*	*F*	Sum
Science	12	36	34	8	90
Humanities	10	24	46	10	90
Business	8	30	70	12	120
Sum	30	90	150	30	300

uniquely cross-classified according to a pair of attributes. To illustrate, Table 12.15 contains information on a sample of 300 students who are classified by major and by grade earned in a basic statistics course.

This type of table, which has one basis of classification across the columns (in this case, grade) and another across the rows (major), is known as a *contingency table*. Because Table 12.15 has three rows and four columns, it is called a three-by-four (often written 3×4) contingency table. In general notation, in an $r \times c$ contingency table (see Table 12.16), where r denotes the number of rows and c the number of columns, there are $r \times c$ cells.

We want to find out whether these data provide strong enough evidence to support the hypothesis that the majors and the grades are somehow related. To solve this problem, we need to compare the observed frequencies with the expected frequencies. When the expected frequencies are far away from what we observed, the test statistic yields a large value that leads to rejection of the null hypothesis. To compute the test statistic, we must find the expected frequencies.

First, we note that of the 300 students surveyed, 90 are science majors, 90 are humanities majors, and 120 are business majors. That means the distribution of students among the three majors is 30 %, 30 %, and 40 %, respectively.

If the students' majors are independent of their performance, the distribution of grades among the 3 majors should also be 30 %, 30 %, and 40 %, respectively. In other words, because science majors make up 30 % of the population, they would be expected to receive 30 % of each grade. That means we expect science majors to account for 9 of the 30 As, 27 of the 90 Bs, 45 of the 150 Cs, and 9 of the 30 Fs. Similarly, we can obtain the expected frequencies for humanities and business majors. This process of obtaining expected frequencies is summarized in Table 12.17. From Tables 12.15 and 12.17, we can calculate the chi-square statistic of Eq. 12.19 as indicated in Table 12.18. The chi-square statistic is $\chi^2 = 11.37$.

The degrees of freedom in this question can be obtained by the formula

$$(r-1)(c-1) = (3-1)(4-1) = 6$$

where r is the number of rows (majors) and c is the number of columns (grades). Note that once $(r–1)(c–1)$ cells are known, the remaining cells are determined (if marginals are known). From Table A5, we find $\chi^2_{6,.05} = 12.59$. Because $12.59 > 11.37$, we accept the null hypothesis that the majors and the grades received are independent of each other.

Table 12.16 Cross-classification of n observations in an $r \times c$ contingency table

	Attribute B				
Attribute A	*1*	*2*	...	*c*	Totals
1	0_{11}	0_{12}	...	0_{1c}	RS_1
2	0_{21}	0_{22}	...	0_{2c}	RS_2
⋮	⋮	⋮	...	⋮	⋮
r	0_{r1}	0_{r2}	...	0_{rc}	RS_r
Totals	CS_1	CS_2		CS_c	n

Table 12.17 Expected grade frequency distribution of 300 students

	A	B	C	F	Sum
Science	9	27	45	9	90
Humanities	9	27	45	9	90
Business	12	36	60	12	120
Sum	30	90	150	30	300

Table 12.18 Computation of chi-square test statistics for student performance in different majors

f_i^o	f_i^e	$(f_i^o - f_i^e)^2/f_i^e$
12	9	1
36	27	3.00
34	45	2.69
8	9	.11
10	9	.11
24	27	.33
46	45	.02
10	9	.11
8	12	1.33
30	36	1
70	60	1.67
12	12	0
Sum 300	300	11.37

A MINITAB solution to this example is presented in Fig. 12.5. The format of this table is similar to that of Table 12.18 with the observed value f_i^o and the expected value f_i^o in each cell.

12.8 Business Applications

In this section, we use 6 examples to show how ANOVA and the χ^2-test can be used to make business decisions.

Application 12.1 Comparing Cash Compensation for Different Groups of Corporate Executives. *Business Week's* Executive Compensation Scoreboard is *BW's* annual report of the total cash compensation (salary and bonus) of the top corporate executives. The table lists the data from the 1986 report (*Business Week,*

```
MTB > READ INTO C1-C4
DATA> 12 36 34 8
DATA> 10 24 46 10
DATA> 8 30 70 12
DATA> END
      3 rows read.
MTB > CHISQUARE USING C1-C4

Chi-Square Test

Expected counts are printed below observed counts

              C1        C2        C3        C4     Total
     1        12        36        34         8        90
           9.00     27.00     45.00      9.00

     2        10        24        46        10        90
           9.00     27.00     45.00      9.00

     3         8        30        70        12       120
          12.00     36.00     60.00     12.00

Total         30        90       150        30       300

Chi-Sq =  1.000 +   3.000 +   2.689 +   0.111 +
          0.111 +   0.333 +   0.022 +   0.111 +
          1.333 +   1.000 +   1.667 +   0.000 =   11.378
DF = 6, P-Value = 0.077
```

Fig. 12.5 MINITAB output for test for independence for student grade distribution

May 4, 1987, pp. 59–94). Assume that the data represent independent random samples of the 1986 total cash compensations for 8 corporate executives in each of 3 industries—banks, utilities, and office equipment/computers. Also assume that the experiment is a completely randomized design.

1986 Total cash compensation for three groups of executives (thousands of dollars)

Banks and bank Holding companies	Utilities	Office equipment and computers
$ 755	$520	$438
712	295	828
845	553	622
985	950	453
1,300	930	562
1,143	428	348
733	510	405
1,189	864	938

To test whether there is a difference in the 1986 total cash compensation for the 3 groups of corporate executives, we use SAS to generate the following ANOVA table.

SAS 22:22 Sunday, March 15, 1992 5

Analysis of Variance Procedure

Dependent variable: COMPENS

Source		DF	Sum of squares	Mean square	F-value	Pr > F
Model		2	685129.3333	342564.6667	6.45	0.0065
Error		21	1115232.5000	53106.3095		
Corrected	Total	23	1B00361.8333			
		R-square	CV	Root MSE	COMPENS mean	
		0.380551	31.95859	230.4481	721.083333	

SAS 22:22 Sunday, March 15, 1992 6

Analysis of Variance Procedure

Dependent variable: COMPENS

Source	DF	ANOVA SS	Mean Square	F-value	Pr > F
Group	2	685129.3333	342564.6667	6.45	0.0065

From Table A6 of Appendix A, we found $F_{.01,2,21} = 5.78$. This value is smaller than 6.450 indicated in the ANOVA table. Therefore, we cannot accept the hypothesis that the 1986 total cash compensation for the 3 groups of corporate executives is equal at $\alpha = .01$.

Application 12.2 Effects of Visual Display Scale on Estimates of Duration. Professor Bobko et al. investigated the effects of visual display scale on duration estimates.[3] They solicited 72 undergraduate volunteers (36 females, 36 males) from an introductory course in psychology. The experimental stimuli were 3 commercially available black-and-white television sets with 3 different screen sizes. The screens had diagonal measurements of .13 m (small), .28 m (medium), and .58 m (large). Using ANOVA with interaction to analyze the empirical data, these researchers got the results listed in Table 12.19.

This value is larger than the critical value $F_{2,66,.05} = 3.11$ (obtained by interpolation from Table A5 in Appendix A), and the p-value for the factor is .005, so the null hypothesis that the display scale does not affect the duration estimates should be rejected. The effect of sex was marginally significant at $F_{1,66} = 3.73$, p-value $= .06$. The interaction of screen size and sex was not significant.

[3] D. J. Bobko, P. Bobko, and M. A. Davis (1986), "Effects of Visual Display Scales on Duration Estimates," *Human Factor* 28, 153–158. Reprinted with permission. Copyright © 1986 by The Human Factors Society, Inc. All rights reserved. The duration is estimated by the length of time passing as a moving display is watched.

Table 12.19 Analysis of variance for the effects of screen size and sex on estimates of duration

Source	Sum of squares	Df	Mean square	F ratio
Screen size	12.73	2	6.36	5.81[a]
Sex	4.08	1	4.08	3.73[b]
Interaction	.29	2	.15	.13
Residual	72.26	66	1.09	
Total	89.36	71		

[a] $p < .005$;
[b] $p < .06$

Application 12.3 Current Ratios for Failed and Nonfailed Firms. In our example from Sect. 12.2, involving starting salaries of economics graduates, each of the 3 treatments consisted of the same number of sample observations. Though it may be more convenient to work with samples of equal size, it is not always possible.

We apply the technique of one-way analysis of variance where the sample observations are not of equal size. In this application, we will test whether the population mean current ratio for two classifications of firms, failed and nonfailed, is significantly different.

Table 12.20 contains the sample current ratios for 6 nonfailed firms and 8 failed firms. The table also includes the sample mean for each treatment and a global mean. Our purpose is to test the following null hypothesis against the following alternative:

H_0: The mean current ratio for nonfailed firms
= the mean current ratio for failed firms.
H_1: The mean current ratios for nonfailed and for failed firms are not equal.

The overall mean is calculated from the data as follows:

$$\bar{x} = \frac{(2.267)(6) + (1.6125)(8)}{14} = 1.892871$$

First, the within-group sum of squares (SSW) is calculated; it is presented in Table 12.21. Accordingly, SSW = 1.1736 + .3891 = 1.5627.

Next we pursue a measure of between-groups variability. In this example, the between-groups variability (SST) is determined as follows:

$$\text{SST}_1 = (2.267 - 1.8928571)^2 = .1400$$
$$\text{SST}_2 = (1.6125 - 1.8928571)^2 = .0786$$

Therefore, SSB = 6(.1400) + 8(.0786) = 1.4686983. Note that each squared discrepancy is weighted by the number of observations in each treatment.

Finally, we calculate the total sum of squares of the two treatments.

$$\text{TSS} = 1.5627 + 1.4686983 = 3.0313983$$

Table 12.20 Current ratios for nonfailed and failed firms

Nonfailed	Failed
2.0	1.8
1.8	1.9
2.3	1.7
3.1	1.5
1.9	1.2
2.5	1.8
	1.6
	1.4
$n_1 = 6$	$n_2 = 8$
$\bar{x}_1 = 2.267$	$\bar{x}_2 = 1.6125$

Table 12.21 Within-group sum of squares

Nonfailed	Failed
$(2.0-2.267)^2$	$(1.8-1.6125)^2$
$(1.8-2.267)^2$	$(1.9-1.6125)^2$
$(2.3-2.267)^2$	$(1.7-1.6125)^2$
$(3.1-2.267)^2$	$(1.5-1.6125)^2$
$(1.9-2.267)^2$	$(1.2-1.6125)^2$
$(2.5-2.267)^2$	$(1.8-1.6125)^2$
	$(1.6-1.6125)^2$
	$(1.4-1.6125)^2$
$SSW_1 = 1.1736$	$SSW_2 = .3891$

In order to test our hypothesis, we must calculate an unbiased estimate of the within-group and between-groups variances. Again, we find the estimate of the within-group variance by dividing the total sum-of-squares deviations of the within-groups variability (SSW) by the appropriate degrees of freedom (n–J).

$$\text{MSW} = \frac{1.5627}{12} = .130225$$

Similarly, we find the estimate of the between-groups variance by dividing the total sum-of-squares deviations of the between-groups variability (SST) by the appropriate degrees of freedom (J–1).

$$\text{MST} = \frac{1.4686983}{1} = 1.4686983$$

Our calculated value of the F ratio is

$$F = \frac{\text{MST}}{\text{MSW}} = \frac{1.4686983}{.130225} = 11.27816$$

From the F distribution in Table A6 with $(J–1)$ and $(n–J)$ degrees of freedom and a .05 significance level, the critical value is 4.75. Because the calculated F ratio is greater than the critical value, we do not accept the null hypothesis that the mean current ratios of failed and nonfailed firms are equal.

Application 12.4 Distribution of Stock Rates of Return. In financial analysis, we are often interested in whether the rate of return of a certain stock follows a normal distribution. The example that follows demonstrates how we used the goodness-of-fit test to find out whether the rates of return of a mutual fund follow a normal distribution.

A stock analyst collected the annualized daily rates of return x_i of a mutual fund in the past 200 trading days and got a mean $\bar{x}$ of 15 % and a standard deviation s_x of 5 %. The rates of return are summarized in Table 12.22. Do the data support rejecting the hypothesis that the rates of return follow a normal distribution? A test at 5 % level of significance follows.

To do the test, we first formulate the hypotheses.

H_0: The annualized daily average mutual fund rates of return are normally distributed with a mean of 15 % and a standard deviation of 5 %.

H_1: The annualized daily average mutual fund rates of return are not normally distributed with a mean of 15 % and a standard deviation of 5 %.

Second, we need to calculate the theoretical frequency (f_i^o) in accordance with the standard normal distribution table (Table A3 in Appendix A). The computation procedure is presented in Table 12.23.

Finally, we calculate χ^2 in terms of Eq. 12.19, as indicated in Table 12.24. The test statistic $\chi^2 = 236.69$. If a level of significance of $\alpha = .05$ is selected, the critical value of χ^2 with 2 degree of freedom is 5.991. Because $281.92 > 5.991$, we conclude that the annualized daily mutual fund rates of return are not normally distributed with mean 15 % and standard deviation 5 %.

Application 12.5 Market-Share Pattern of a New Cereal Product. G. A. Churchill proposed a goodness-of-fit technique to test the market-share pattern of a new cereal called score produced by a breakfast food manufacturer.[4] The cereal was packaged in three standard sizes: small, large, and family size. The manufacturer's experience with other cereals suggested that, for every small package, three of the large and two of the family size are also sold. The manufacturer wanted to know whether this same consumption pattern would hold with score, because a change in consumption pattern could have significant implications for production and packaging. The manufacturer therefore decided to conduct a market test over a 1-week period. In this period, 1,200 boxes of the new cereal were sold. The distribution of sales by size is given in Table 12.25.

[4] G. A. Churchill, Jr. (1983), *Marketing Research: Methodological Foundations,* 3rd. ed., pp. 523–524. Copyright © 1983 by The Dryden Press, reprinted by permission of the publisher.

Table 12.22 Rate-of-return data for 200 days

Rates of return, x_i (%)	Observed frequency, f_i^o
Under − 5	20
−5 to under 0	33
0 to under 10	48
10 to under 20	41
20 to under 30	29
30 or above	29
Total	200

Table 12.23 Computation of theoretical frequencies in each rate-of-return interval

Class boundaries	x	$z = (x–15)/5$	Area under standard normal curve left of x	Area of class interval (P)	Expected frequency if H_0 is true, $f_i^o = 200P$	
Under −5	−5	−4	0	0	0	31.74
−5 to under 0	0	−3	.0013	.0013	.26	
0 to under 10	10	−1	.1587	.1574	31.48	
10 to under 20	20	1	.8413	.6826	136.52	31.74
20 to under 30	30	3	.9987	.1574	31.48	
30 and above	∞	∞	1.0000	.0013	26	
Total				1.0000		

Table 12.24 Worksheet for computing the test statistic χ^2

Class boundaries	f_i^o	f_i^e	$(f_i^o - f_i^e)$	$(f_i^o - f_i^e)^2/f_i^e$
Under 10	101	31.74	69.26	151.13
10 to under 20	41	136.52	−95.52	66.83
20 and above	58	31.74	26.26	21.73
Total	200	200	0	239.69

Table 12.25 Distribution of boxes of new cereal sold

Small	Large	Family	Total
240	575	385	1,200

To test whether the relative frequencies of the various sizes of the new product are the same as those of the old product, we can test the hypotheses

$$H_0 : P_1 = \frac{1}{1+3+2} = \frac{1}{6}, \quad P_2 = \frac{3}{1+3+2} = \frac{1}{2}, \quad P_3 = \frac{2}{1+3+2} = \frac{1}{3}$$

H_1 : At least one of these P_is is incorrect.

To perform this test, we first calculate the expected frequencies:

$$f_i^e = \frac{1,200}{6} = 200, \quad f_2^e = \frac{1,200}{2} = 600, \quad f_2^e = \frac{1,200}{3} = 400$$

Substituting both expected and observed frequencies into Eq. 12.19, we obtain

$$\chi^2_{3-1} = \frac{(240-200)^2}{200} + \frac{(575-600)^2}{600} + \frac{(385-400)^2}{400} = 9.60$$

From Table A5, we find that $\chi^2_{2,.05} = 5.99$, which is smaller than 9.60. Hence, we reject H_0 and accept H_1. In other words, the null hypothesis of sales in the ratio of 1:3:2 is rejected. This result suggests that the sale of the new cereal, score, will follow a different pattern.

Application 12.6 The Effect of Price Advertising on Alcoholic Beverage Sales. To study the effect of price advertising on alcoholic beverage sales, G. B. Wilcox examined the effects of price advertising on sales of beer in Lower Michigan. Wilcox used Michigan in his study because since 1975, except for the short period from March 1982 until May 1983, Michigan has banned retailers from advertising the price of beer products. The data he used covers 3 different periods and are presented in Table 12.26.

To examine whether there is sufficient evidence to indicate differences in the average total sales of beer in the 3 periods, we use SAS to generate the following ANOVA output.

SAS 23:34 Sunday, March 15, 1992

Analysis of Variance Procedure Dependent variable: Sales

Source	DF	Sum of squares	Mean square	F-value	Pr > F
Model	2	11357.82500	5678.91250	0.78	0.4760
Error	15	109152.67500	7276.84500		
Corrected Total	17	120510.50000			

R-square	C.V.	Root MSE	Sales mean
0.094248	16.39944	85.30443	520.166667

SAS 23:34 Sunday, March 15, 1992 3

Analysis of Variance

Procedure Dependent variable: SALES

Source	DF	Anova SS	Mean square	F-value	Pr > F
PERIOD	2	11357.82500	5678.91250	0.78	0.4760

From Table A6 in Appendix A, we found $F_{.05,2,15} = 3.68$. This number is larger than .78 indicated in the above ANOVA table; therefore, we cannot reject the hypothesis that price advertisements on alcoholic beverages did not affect sales of beer in the three different periods in the state of Michigan.

Table 12.26 Bimonthly beer sales for three different periods (units: thousands of 31-gal barrels)

Period 1: Price advertising restricted (May/June 1981–Jan./Feb.1982)	Period 2: No restrictions (March/April 1982–May/June 1983)	Period 3: Price advertising restricted (July/August 1983–March/April 1984)
462	522	433
417	508	470
516	427	609
605	477	442
654	603	446
	692	
	584	
	496	

Source: G. B. Wilcox, "The Effect of Price Advertising on Alcoholic Beverage Sales," *Journal of Advertising Research,* Vol. 25, No. 5, October/November 1985, 33–37

12.9 Summary

Using the basic concepts of mean, variance, and F statistics discussed in previous chapters, we explored a statistical method called analysis of variance for testing the difference between sample means. We also examined the use of the chi-square statistic in goodness-of-fit tests and in testing the assumption of independence. Several applications of analysis of variance in business decisions were discussed in some detail.

Questions and Problems

1. The following table shows the sales figures for 4 salespeople on 3 randomly selected days. Use analysis of variance to test the hypothesis that the mean daily sales figures (in thousands of dollars) are the same for all 4 salespeople. That is, test H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. Use $\alpha = .05$.

	Salesperson			
	1	2	3	4
Day	15	9	17	12
	17	12	20	13
	22	15	23	17
Mean	18	12	20	14
Variance	13	9	9	7

2. The yearly portfolio returns for 3 different investment firms over a 10-year period are listed in the accompanying table. Do these data show a statistically significant difference in the firms' performances? Assume that the population errors meet the conditions necessary for ANOVA. Let $\alpha = .01$.

	Firm 1	Firm 2	Firm 3
Mean return	11.5	12.0	10.0
Standard deviation	3.0	2.0	2.4
Number of investments	20	10	25

3. A consumer organization wants to compare the prices charged for a particular dishwasher in 3 types of stores in a suburban county: discount stores, department stores, and appliance stores. Random samples of 4 discount stores, 6 department stores, and 5 appliance stores were selected. The results were as shown.

Discount	Department	Appliance
12	15	15
14	18	18
16	18	16
15	14	16
	18	19
	15	

At the .05 level of significance, is there any evidence of a difference in the average price between the types of stores? Use the MINITAB commands presented in Fig. 12.2 to answer this question.

4. Three packaging materials were tested for moisture retention by storing the same food product in each for a fixed period of time and then determining the moisture loss. Each material was used to wrap 10 food samples. The results are given in the accompanying table.

   (a) Construct the ANOVA table.
   (b) Can we reject the hypothesis that the materials are equally effective? Use $\alpha = .05$.

	Material 1	Material 2	Material 3
Number of packages	10	10	10
Mean loss	231	238	224
Sample variance	40	38	30

5. The quality control manager at a sugar refinery was worried that two packaging production lines might be filling the packages with different weights. Samples of size 16 were taken from each of the production lines, and the contents of these 10-lb packages were carefully weighed. The following sample results seem to indicate that the mean weights of the 10-lb packages from the two lines

are the same, but there appears to be much more variation in the weights of the packages coming off the second line. Test the hypothesis that the two lines have the same variation in weights by testing H_0: $\sigma_1^2 = \sigma_2^2$ with $\alpha = .05$. Use H_1: $\sigma_1^2 < \sigma_2^2$. If the following data do not support rejection of the null hypothesis, what must we do to test H_0: $\mu_1 = \mu_2$?

$$\begin{aligned} n_1 &= 16 & n_2 &= 16 \\ \overline{X}_1 &= 10.15 & \overline{X}_2 &= 10.16 \\ S_1 &= .07 & S_2 &= .1 \end{aligned}$$

6. The Environmental Protection Agency is studying coliform bacteria counts at the beaches of a large suburban county. Three types of beaches are to be considered (ocean, bay, and sound) in three geographic areas of the county (west, central, and east). Two beaches of each type are randomly selected in each region of the county. The coliform bacteria counts (in parts per thousand) at each beach on a particular day were as shown in the table.

	Geographic area		
Type of beach	West	Central	East
Ocean	25	9	3
	20	6	6
Bay	32	18	9
	39	24	13
Sound	27	16	5
	30	21	7

   (a) Use the .05 level of significance and determine (1) whether there is an effect due to type of beach, (2) whether there is an effect due to geographic area, and (3) whether interaction between type of beach and geographic area has an effect.
   (b) What conclusions about average bacteria count do the results support?

7. Explain the difference between the one-factor and the two-factor ANOVA models.
8. A researcher has an ANOVA problem with five columns (treatments). Would you recommend testing for differences between the columns by looking at pairs of columns? Explain why. How many pairs are there?
9. A researcher concludes that there is no difference in the column treatments in a one-factor model. Upon reanalysis of the same data via two-factor ANOVA, however, the researcher concludes that there is a difference in the column treatments. Did the researcher make a mistake? Which conclusion (if either of them) is true? Explain how these contradictory conclusions can (or cannot) be justified.

10. A secretarial training school is experimenting with four different manuals for a typing course. The school divided 20 students into four classes, and each class used a different manual. At the end of the training session, a test was given, and the scores shown in the table were reported. Do the data support the hypothesis that the four different manuals create different effects? Do a 5 % test.

Manual

A	B	C	D
78	67	75	89
74	79	95	86
95	85	69	87
93	79	60	87
85	86	94	73

11. A research institution says that gasoline is gasoline. That is, there is no difference among different brands of gasoline in terms of mileage per gallon. An independent consumer rights organization did an experiment on three different brands of gasoline. It divided cars of the same make and the same condition into three groups, and the members of each group were filled with a different brand of gasoline. The test results follow. Do the data support the hypothesis of no difference among gasolines? Do a 5 % test.

Gasoline		
A	B	C
34	29	32
28	32	34
29	31	30
37	43	42
42	31	32
27	29	33
29	28	

12. A college professor taught an interdisciplinary course in the last 3 years to students of different majors. He believes that the students' majors do not have any impact on their performance in the class. He picked 24 students who represented 3 different majors and recorded their test scores. Do the data necessitate rejection of the professor's hypothesis? Do a 5 % test.

Business	Humanities	Science
85	84	63
57	95	73
92	87	83
83	73	64
84	83	79
83	73	74
75	85	65
73	72	98

13. A doctor recruited 15 volunteers and put them on 3 different diets that were supposed to lower the subjects' cholesterol levels. The effects of the diet plans in terms of lowered cholesterol level are recorded in the table. Do the data cause as to reject the null hypothesis that the three different diet plans have the same effect? Do a 5 % test.

Diet plan		
A	B	C
20	12	13
14	15	13
21	21	23
15	16	19
17	18	14

14. General Motors, the largest auto producer in this country, produces and sells its cars under five different brand names. Some of the cars sold under different names can be considered "sister cars" that should turn in the same performance. An auto analyst wants to see whether the "sister cars" sold under different names do indeed have the same performance. He tested 20 cars from three different brands and recorded the mileages per gallon. Do the data require rejection of the null hypothesis that the mileages per gallon generated by three "sister cars" are the same? Do a 5 % test.

Brand		
A	B	C
32	31	34
29	28	25
32	30	31
25	34	37
35	39	32
33	36	
34	38	
31		

15. Market analysts want to investigate the popularity of three types of radio programs. Their market research yielded the following ratings over the last 5 months. Can the analysts conclude that the three types of programs have different ratings? Do a 5 % test.

Time	Talk show	Sports show	Music
Early morning	20	30	17
Late morning	17	15	15
Afternoon	18	21	12
Evening	23	27	18
Midnight	25	22	11

16. In question 15, the market analyst hired a statistician to do the research. The statistician, realizing that these programs are broadcast at five different times of day, took the time factor into consideration. Do a test for the same hypothesis. Use 5 %.
17. On a beach, there are three ice cream stands that are supposed to be occupying equally good locations. The management of the beach wants to know whether this assumption is true. The ice cream sales during the past few days, in hundreds of dollars, are recorded here. Do a 5 % test to determine whether the ice cream sales are the same in the three locations.

Location		
A	B	C
12	31	21
14	21	14
14	20	17
18	17	16
21	12	23

18. A stock analyst thinks four stock mutual funds generate about the same return. She collected the accompanying rate-of-return data on four different mutual funds during the last 5 years.

	Fund			
	A	B	C	D
1988	12	11	13	15
1989	12	17	19	11
1990	13	18	15	12
1991	18	20	25	11
1992	12	19	19	10

(a) Do a one-way ANOVA to decide whether the funds give different performances. Use 5 %.
(b) Do a two-way ANOVA to decide whether the funds give different performances. Use 5 %.

19. The personnel office recently produced a set of aptitude test questions designed to determine whether a potential employee can be a good salesperson. The test was tried on current employees before it was used for future employees. The test scores on employees in three different departments are listed here. Do the tests generate different scores for different departments? Do a 5 % test.

Department		
Accounting	Sales	Production
78	85	76
79	87	77
80	89	97
72	79	71
87	99	81
98	95	79

Use the following information to answer questions 20–23. An investor recorded the performance of stock mutual funds during the last 3 years. He classified the stock mutual funds into three categories, growth, income, and mixed. The rates of return during the last 3 years are presented here:

	Type		
Year	Growth	Income	Mix
1990	12	14	15
	17	12	16
	19	12	17
	17	13	15
1991	18	19	18
	21	14	15
	21	16	17
1992	22	13	15
	21	15	18

20. Use MINITAB to do a one-way ANOVA to determine whether the 3 different years have the same average rate of return. Use the 5 % level of significance.
21. Use MINITAB to do a one-way ANOVA to determine whether the 3 different types of mutual funds have the same rate of return. Use the 5 % level of significance.
22. Use MINITAB to do a two-way ANOVA to determine whether the three different types of mutual funds have the same rate of return. Use the 5 % level of significance. (Hint: Follow the procedure presented in Fig. 12.4.)
23. Use MINITAB to do a two-way ANOVA to determine whether the three different years have the same average rate of return. Use the 5 % level of significance. (Hint: Follow the procedure presented in Fig. 12.4.)
24. A researcher contacted 1,000 doctors and asked them what kind of pain reliever they would like to have with them if they were stranded on a desert island. The responses were

Brand A	Brand B	Brand C	Others
250	230	260	260

Do a test to determine whether the 4 categories of products receive the same number of recommendations from doctors. Use 5 %.

25. Four instructors teach introductory-level economics courses during the same time period. The numbers of students taking their courses are

Instructors			
A	B	C	D
100	120	90	110

Can we reject that the four instructors are about equally popular among the students? Do a 5 % test.

26. The personnel manager wants to know whether an equal number of employees call in sick on the 5 days of the regular work week. The sick days recorded during last year were distributed as follows:

M	T	W	Th	F
40	30	32	25	45

Can we conclude that the five different weekdays have different frequencies of sick calls? Do a 5 % test.

27. An economics consulting company wants to study bank managers' opinions about what lending rate will prevail for the next 3 months. It sends questionnaires to 940 bank managers and gets the following responses:

Higher	Lower	Same	No idea
210	220	210	300

(a) Can we reject that all of the four opinions are held by about the same number of bank managers? Do a 5 % test.
(b) Can we reject that the three kinds of opinions (excluding "No idea") are held by about the same number of bank managers? Do a 5 % test.

28. A management consulting company is interested in how managers look at the prospects for the economy in this country. Questionnaires were sent to 2,000 managers in different areas of the country. The responses were:

	Future prospects		
	Optimistic	Pessimistic	No change
Top-level managers	300	250	100
Middle managers	200	200	220

(a) Does the evidence suggest that middle managers are equally split among the three different opinions? Do a 5 % test.
(b) Does the evidence suggest that middle management's opinion pattern is similar to that of top-level management? Do a 5 % test.

29. Lotteries are getting more and more popular in this country. Many books claiming to teach people how to pick winning numbers have been published. According to an official in the state's lottery office, the game is fair in the sense that even number has the same chance of being drawn. Briefly explain how you can test this contention. (And remember that if it is true, any money spent on "systems" or "secrets" for winning the lottery is money wasted.)
30. Professor Maloy uses different textbooks to teach statistics to two different college classes. Book 1 is the standard textbook also used by other instructors. Book 2 is a more recently introduced text. Over the years, Maloy recorded the grade distribution of the students.

	A	B	C	D	F
Book 1	20	80	100	20	10
Book 2	5	22	30	4	1

Can the professor conclude that the grade distribution pattern of students using book 1 is different from the grade distribution pattern of students using book 2?

31. An insurance company reviewed its policyholders' records during last year and organized the data in the following table. Do the data dictate rejection of the null hypothesis that the accident pattern has a Poisson distribution? Do a 5 % test.

Number of accidents	Number of policyholders
0	2000
1	150
2	10
3	1

32. The number of patients to arrive at an emergency room each day is recorded in the accompanying table. The average number of emergency room patients is approximately 10. Do the data support the hypothesis that the number of emergency room patients follows a Poisson distribution? Do a 5 % test.

Number of patients	0	1	2
Number of days	400	14	1

33. An eight o'clock train that pulls into Penn Station in New York City every weekday has a 20 % chance of being late. A supervisor from the Port Authority recorded the number of days that the train arrived late each week for the last 100 weeks. Does the evidence compel us to refute the null hypothesis that the data come from a binomial distribution? Do a 5 % test.

Number of late arrivals in a week	0	1	2	3	4	5
Number of weeks	5	15	30	30	15	5

34. A nationwide testing service collected scores from a set of examination questions that were used during the last 3 years. The distribution is summarized in the accompanying table. The mean is 600 and the standard derivation is 200. Does the evidence support rejection of the hypothesis that the data come from a normal distribution? Do a 5 % test.

Score	Frequency
<300	10
301–400	25
401–500	40
501–600	45
601–700	44
701–800	35
801–900	20
More than 900	15

35. The daily rate of return for a stock (adjusted to an annual rate) is summarized in the following table. Can you show that these data do not come from a normal distribution? The mean is 0 % and the standard derivation is 1.6 %. Do a 5 % test.

Rate of return	Frequency
Less than −3 %	20
−3 % to −2 %	25
−2 % to −1 %	30
–l % to 0	50
0–1 %	40
1–2 %	25
More than 2 %	5

36. The highway bureau records the following numbers of accidents during the past 365 days. Does this frequency distribution cause us to reject the null hypothesis that the accidents exhibit a Poisson distribution? Do a 5 % test.

Number of accidents in a day	Number of days
0	320
1	30
2	10
3	5

37. A college professor wants to use a normal distribution to analyze his students' grades. He randomly selects 200 previous grades and organizes them in the following table. The mean is 67.75 and the standard deviation is 13. Does the frequency distribution support the hypothesis that students' grades follow a normal distribution? Do a 5 % test.

Grades	Frequency
≤ 40	2
41–50	15
51–60	35
61–70	70
71–80	40
81–90	28
>90	10

38. In a hospital, 100 patients were checked for their cholesterol level. The mean was 200 and the standard deviation was 27.2. Do the data collected support the hypothesis that the cholesterol levels follow a normal distribution? Do a 5 % test.

Cholesterol Level	Frequency
≤ 160	4
161–180	21
181–200	25
201–220	30
221–240	10
241–260	8
>260	2

39. A college professor has taught business statistics for the last 5 years. He used standard tests every semester. The distribution of grades during the past 5 years was

A	B	C	D	F
15 %	30 %	40 %	10 %	5 %

This professor has just finished grading students this semester and has found that the frequency distribution of the grades is

A	B	C	D	F
6	12	15	5	1

He feels that this semester's students have a different grade distribution pattern. Do you agree with him? Do a 5 % test.

40. A travel agency was curious about whether the service a guest receives is related to the size of the hotel. The agency surveyed 300 customers and summarized their responses in the accompanying table. Determine whether the data support the hypothesis that the customer's opinion and the size of the hotel are related. Each customer gave only one opinion for one size of hotel. Use 5 %.

	Size of hotel		
	Large	Midsize	Small
Satisfied	80	40	30
So-so	60	30	10
Dissatisfied	20	20	10

Use the following information to answer questions 41–43. Four hundred and fifty economists of different ideologies were asked to forecast the prospects for the economy during the Bush administration. Here's what they said:

	Opinion		
Ideology	Boom	So-so	Recession
Conservative	80	60	60
Liberal	60	40	40
Radical	50	30	30

41. Do the data support the hypothesis that ideology and opinion are related? Do a 5 % test.
42. Test, at the 5 % level, the hypothesis that about equal numbers of economists hold each opinion.
43. Can you say that the opinion pattern of the liberal economists is the same as the opinion pattern of the radical economists? Do a 5 % test.
44. A magazine is interested in knowing whether which newspaper is read and level of education of the reader are related.

	Newspaper		
Education	Post	News	Tribune
High school	300	200	100
College	200	300	100
Graduate school	100	200	300

Use MINITAB to determine whether newspaper read and education are related. Use the 5 % level of significance. (Hint: Follow the procedures presented in Fig. 12.5.)

45. The sales manager wants to know whether salespeople's performance is related to their zodiac sign. Three hundred salespeople were surveyed. Their performance is summarized in the following table.

	Performance		
Zodiac sign	Good	Mediocre	Bad
Leo	80	30	20
Gemini	50	20	10
Virgo	40	40	10

Do the data support the belief that the performance and zodiac sign of a salesperson are related? Do a 5 % test.

46. An advertising agency wants to know whether there is a relationship between TV shows and the age of the audience. The following data were compiled.

	Age of the audience			
TV show	10 and younger	Teenager	20–40	40 and older
Game show	100	120	200	400
Sitcom	20	120	400	200
News	2	40	48	50

Do the data support the hypothesis that the age of the audience is related to the type of show that is preferred? Do a 5 % test.

Use the following information to answer questions 47–49. A developer asks visitors how they heard of the housing project they are looking at. Their responses are shown in the following table.

Distance from construction site	*Source of information*			
	Referred by a friend	Newspaper	Radio	TV
Within 10 miles	40	200	120	150
10–30 miles	30	180	120	120
Farther than 30 miles	10	150	100	100

47. Is the distance from the construction sites independent of the way people hear of the housing project? Do a 5 % test.
48. Can we conclude that the effects of publicizing the project in the 3 different media (newspaper, radio, and TV) are different? Do a 5 % test.
49. If you live 10–30 miles away from the construction site, is your sources of information pattern the same as that of the people living within 10 miles? Do a test of 5 %.

Use the following information to answer questions 50–55. A company operates three mutual funds. The managers of these mutual funds invest the money entrusted to them in different kinds of assets. The rates of return in the last 5 years are recorded in the following table.

	Real estate fund	Government bond fund	Stock fund
1985	6 %	7 %	3 %
1986	20	8	9
1987	6	12	8
1988	15	9	15
1989	3	10	20

The company also randomly sampled its customers and compiled the following table:

	Real estate fund	Government fund	Stock fund
Retirees	30	80	40
51–65	40	60	50
36–50	80	20	50
20–35	40	40	70

50. Do the three different kinds of mutual funds attract about the same numbers of investors? Do a 5 % test.
51. Can we conclude that the 3 different kinds of mutual funds generate about the same average rate of return? Do a 5 % test, using a one-way ANOVA.
52. Do a two-way ANOVA to determine whether the rates of return for the three kinds of mutual funds are about the same. Do a 5 % test.
53. Determine whether the stock fund is equally popular among the four different age groups. Do a 5 % test.
54. Determine whether the investment pattern of the age group 20–35 is the same as that of the age group 51–65. Do a 5 % test.
55. Are fund preference and age group related? Do a 5 % test.
56. A plant that runs three shifts would like to know whether the three shifts are equal in average productivity. The productivity breakdown is presented in the following table.

Day	Shift 1	Shift 2	Shift 3
Monday	30	40	20
Tuesday	40	50	30
Wednesday	40	40	30
Thursday	40	30	20
Friday	20	30	20

(a) Are average productivities of the three different shifts the same? Do a 5 % test, using a one-way ANOVA.
(b) Are the average productivities of the three different shifts different? Do a 5 % test. Consider the weekday factor in testing this hypothesis.

57. A nationwide real estate brokerage house wants to study the relationship between rent per square foot and size of the property. The data collected are summarized in the accompanying table. Using these data, can we reject the null hypothesis that the average rents per square foot are equal? Do a 5 % test.

Size of the property (in square feet)		
Less than 1,000	1,000 to 2,000	2,000 or more
3	2	3
4	5	6
5	5	7
5	6	7
5	5	7
4	6	6

58. A consultant argues that location, the most important factor in the real estate business, was not considered in the test performed in question 57. He suggests redoing the test by controlling the location factor. Do a 5 % test to see whether the conclusion changes when the data are presented as follows:
Size of the property (in square feet)

	Size of the property (in square feet)		
Location	Less than 1,000	1,000 to 2,000	2,000 or more
Bad	3	2	3
	4	5	6
So-so	5	5	7
	5	6	7
Good	5	5	7
	4	6	6

59. The performances of 250 salespeople in a company are summarized in the following table.

Sales	Frequency
Less than 78	13
78–80	37
81–83	40
84–86	50
87–89	60
90 or more	20

Derive the expected frequencies, assuming that the data are from a normal distribution. Do the data collected support the hypothesis that the sales following a normal distribution? Do a 5 % test.

60. A chicken farm came up with 4 different ways of mixing chicken feeds. The feeds were tested on 20 chickens. The results, given in terms of the chickens' weight, are presented in the accompanying table. Do a 5 % test of the hypothesis that the weights resulting from the different feeds are approximately the same.

Group			
A	B	C	D
4.5	4.2	4.1	4.6
4.4	4.3	4.6	4.2
4.5	4.3	4.2	4.9
4.3	4.2	4.3	4.4
4.9	4.9	4.8	4.7

Use the following information to answer questions 61–64. A survey was sent to 400 students to solicit their opinions about a new rule for using the student centers. Here are the results:

Year	Against	Indifferent	Agree
First year	30	30	40
Sophomore	50	40	30
Junior	20	50	30
Senior	10	30	40
Total	110	150	140

61. Do you think the three different opinions have about the same number of responses? Do a 5 % test.
62. Do you think the three different opinions receive about the same amount of support among first-year students? Do a 5 % test.
63. Do you think opinion pattern and year in school are related? Do a 5 % test.
64. Do you think there are equal amounts of support for the new rule from the four different classes? Do a 5 % test.
65. An insurance company is interested in knowing the relationship between traffic accident claims arid the type of cars that policyholders drive. The numbers of accidents per 1,000 automobiles during last year in six states are reported in the accompanying table. Determine whether the three kinds of cars have the same average accident rate. Use a 5 % level of significance.

	Type of automobile		
State	Sports Car	Sedan	Wagon
New Jersey	30	15	16
New York	20	15	17
Connecticut	15	12	11
Massachusetts	17	13	12
Vermont	18	21	15
New Hampshire	17	12	13

66. The accompanying table shows highway patrol data on the numbers of speeding tickets given in the last 3 months. Do the data show that all 3 months have about the same number of tickets? Do a 5 % test.

	Sports cars	Sedans	Wagons	Trucks
April	44	30	32	18
May	46	32	30	25
June	45	35	37	27

67. In a poll, people were asked their opinions about the death penalty. The breakdown of the responses is given in the accompanying table. Do the data show that educational level and opinion are independent of each other? Do a 5 % test.

Educational level	Favor	Oppose
Elementary school	400	200
High school	200	400
College	200	400

68. In a recent survey, people were asked whether they are happy with the current income tax structure. Do the results that follow support the hypothesis that how people feel about the tax structure depends on what tax bracket they are in? Do a 5 % test.

	Satisfied	Dissatisfied	Very dissatisfied
Low bracket	40	40	50
Middle bracket	50	30	30
High bracket	30	50	60

Use the following information to answer questions 69–71. A company puts vending machines in different locations. The numbers of sodas sold (in thousands) in the last 3 months are presented in the following table.

	Location		
	Beach	School gymnasium	Gas stations
April	3	4	4
	3	5	5
	2	5	6
May	6	4	5
	8	5	6
	6	4	7
June	10	4	8
	10	6	7
	12	6	8

69. Do the data support the hypothesis that April, May, and June have the same amount of sales? Do a 5 % test, using one-way ANOVA.
70. Do a one-way ANOVA to see whether you can argue that the three different locations have different amounts of sales. Use 5 %.
71. Do a two-way ANOVA to see whether the three different locations have different amounts of sales. Use 5 %.
72. Hannah, a wine dealer, believes that the taste of wine depends on the year the wine was bottled. Do the data she collected from a recent wine-tasting contest support her belief? Do a 5 % test.

	Year		
	1968	1973	1985
Excellent	40	45	25
Good	35	25	45
Fair	45	15	20
Yuck	10	15	20

73. The manager in a department store believes that whether a customer pays by cash, charge, or check depends on the amount of money spent. Do the following data support what the manager believes? Do a 5 % test.

Amount spent	Charge	Check	Cash
Less than $10	20	30	70
Between $10 and $100	40	40	40
Over $100	60	40	20

74. The demand for different types of automobiles should be related to their owners' needs. A manager in a local auto dealership randomly pulls samples from the dealership's customer files. Do the resulting data support the manager's belief? Do a 5 % test.

	Auto purchased		
	Sedan	Wagon	Sports car
Single	30	5	20
Married, no children	40	15	20
Married, at least one child	30	40	20

75. An advertising agent believes that different types of programs attract audiences of different age groups. She collects the following data to study her claim.

	Program type		
Age group	Sitcom	Game Show	News
10–19	40	40	20
20–29	60	40	50
30–39	60	30	60
40 or older	40	20	40

Determine, at $\alpha = 5\ \%$, whether you can reject her claim.

76. A consumer rights organization wanted to check out different diet plans. It recruited 33 volunteers and sent them to four different programs. After the first 2 weeks, the weight losses, in pounds, were recorded and organized in the accompanying table. Do a 5 % test to determine whether the 4 programs are equally effective.

A	B	C	D
8.0	9.9	8.9	7.6
8.8	9.1	8.2	7.7
8.7	9.8	8.1	7.5
8.6	9.8	8.0	7.8
8.0	9.9	8.6	7.6
8.8	9.6	8.6	7.3
8.5	9.2	8.6	7.1
	9.8	8.4	8.0
			7.5
			8.0

77. The dean of the business school wants to find out whether the instructors in four departments are grading students similarly. The following data are compiled. Do you think the grade distribution depends on the department? Do a 5 % test.

	Finance	Management	Accounting	Marketing
A	35	45	35	25
B	50	60	55	35
C	15	30	25	10
F	30	45	35	20

78. A Consumer Protection Coalition decides to study the delay times, in minutes, for four different airlines: A, B, *C*, and D.

A	B	C	D
25	22	21	30
35	31	24	28
35	33	34	32
30	28	29	27
44	41	40	15
31	32	17	19

It is believed that the average delay times of the 4 airlines are about equal. Do a test at the 5 % level to decide whether the data support rejecting this hypothesis.

79. In question 78, a statistician argues that the length of delay may depend on the airport from which the airplane departs. Accordingly, the data were regrouped to reflect departure sites X, Y, and Z. Here are the results:

	Airlines			
	A	B	C	D
X	25	22	21	30
	35	31	24	28
Y	35	33	34	32
	30	28	29	27
Z	44	41	40	15
	31	32	17	19

Redo the test to decide whether the airlines' delay times are about equal by considering the effect of departure location. Use 5 %.

80. The delay times of 200 delayed flights were compiled in the following frequency distribution. The mean is about 45 and the standard deviation is 20. Do the data follow a normal distribution? Do a 5 % test.

Delay time (in minutes)	Frequency
0–15	20
16–30	32
31–45	48
46–60	52
61–75	38
76–90	10

81. The numbers of missing pieces of luggage are compiled in the following table. Do the data follow a Poisson distribution? Do a 5 % test.

Number of missing Pieces of luggage	Number of flights
0	985
1	10
2	4
3	1
More than 3	0

82. A bank manager is interested in the amount of cash being withdrawn each Friday. He collects data on the last 90 Fridays and compiles them in the accompanying table. The mean is 340 and the standard deviation is 64. Determine whether the amount of cash withdrawn follows a normal distribution. Use a 5 % significance level.

Cash withdrawn	Frequency
Less than 250	5
250–300	21
301–350	25
351–400	20
401–450	15
More than 450	4

83. A financial analyst is interested in conducting an extensive study of credit card debt. He wants to know whether the income of cardholders is related to the size of the debt. He compiles the data in the accompanying table. Determine whether size of debt and income level are independent. Use a 5 % level of significance.

	Size of debt		
Income	$200 to $500	$500 to $1,000	$1,000 and above
Less than $20,000	400	200	100
$20,000 – $40,000	450	500	300
Higher than $40,000	100	200	500

84. There are many books to help people learn to use computer software packages. An instructor checked these books and found that they are all of similar quality. He picked four books and used them in his classes. If the students have the same average grades, he will use the cheapest book. On the basis of the test results that follow, do you think the four classes have about the same grades? Do a 5 % test.

Class			
W	X	Y	Z
43	77	72	72
45	72	73	74
67	75	71	75
68	69	65	65
73	67	68	66
72	66	69	68
55	65	73	74
62	63	72	81

85. It is believed that the quality of a certain product is related to the time of day the product is produced. The following table summarizes the results of tests on some random samples produced in a single day.

Time	Good	So-so	Bad
Morning shift	25	10	5
Afternoon shift	15	20	5
Evening shift	10	20	10

Use a 5 % test to determine whether the quality of the product is independent of the time it was produced.

86. The placement office in a business school randomly sampled 24 graduates from three departments and recorded their starting salaries. Determine whether graduates of the 3 departments have about the same starting salaries. Use a 5 % level of significance.

Management	Marketing	Accounting
$24,550	$25,200	$24,150
24,790	27,200	24,100
24,310	24,100	23,900
24,200	25,400	25,650
24,900	23,300	23,700
25,200	24,200	24,900
23,900	25,000	24,350

87. A magazine wants to know the relationship between people's voting behavior and their level of income. Questionnaires were sent to 200 voters, and FCE responses are summarized here. Do the data support the hypothesis that income and voting behavior are related? Use a 5 % level of significance.

	Income		
	High	Medium	Low
Incumbent	35	22	10
Challenger	25	25	40
Did not vote	10	23	10

Use the following information to answer questions 88–91. A questionnaire was sent to 200 students on the campus, asking them to indicate their ethnic background and give their opinion about race relations on campus. The responses are summarized in the following table.

Ethnic	*Opinion on race relations*		
Background	Good	So-so	Bad
White	40	80	40
African–American	6	8	6
Asian–American	6	10	14

88. Are African–Americans' opinions equally split among the three categories? Do a 5 % test.

89. The makeup of the student body is 75 % white, 15 % African–American, and 10 % Asian–American. If the samples are randomly selected, the samples' ethnic distribution should be similar to the ethnic distribution on the campus as a whole. Do the data support that hypothesis? Do a 5 % test.
90. Are ethnic background and opinion on this issue related? Do a 5 % test.
91. Do the two minority groups have a similar opinion pattern? Do a 5 % test. Assume we know that African–Americans' pattern is exactly 30 % for good, 40 % for so-so, and 30 % for bad.
92. Two hundred and ten people were asked which TV news programs they usually watch. The answers are compiled in the following table. Can you say that the three networks have audiences of about the same size? Do a 5 % test.

Network A	Network B	Network C
80	70	60

93. An election was held in a big city whose population is 50 % white, 40 % black, and 10 % Hispanic. Among the elected, 40 council members are white, 30 are black, and 10 are Hispanic. Do we have enough evidence to say that the three ethnic groups are represented on the council in proportion to their representation in the population?
94. Do workbooks make a difference in students' performance? A statistics instructor uses her class as a sample. Do the results suggest that the grade patterns of those who own a workbook and of those who do not are different? Do a 5 % test.

	Grade				
	A	B	C	D	F
Own a workbook	5	4	6	2	1
Don't own a workbook	2	6	4	3	2

95. The president of a local bank suspects that his employees care only about the big customers. He randomly sampled 325 loans made during the last year and asked the borrowers their opinion of the service they received. On the basis of the results, do you think loan size and service received are independent? Do a 5 % test.

	Loan size		
Service	Small	Midsize	Large
Satisfied	10	20	40
Acceptable	20	45	30
Dissatisfied	33	33	24

96. A gambler wants to know whether the dice used in a casino are fair. If the dice are fair, the probabilities of seeing 1, 2,. . ., 6 are all $\frac{1}{6}$.. The gambler recorded the outcomes of 600 rolls of the dice. Here are his results:

1	2	3	4	5	6
98	93	107	105	97	100

Do the data support the hypothesis that the dice are fair? Do a 5 % test.

97. A magazine wants to study the relationship between people's education and the medium they are exposed to the most. Questionnaires were sent to 100 people of different educational backgrounds. The results are summarized here. Are educational background and medium used the most related? Do a 5 % test.

	TV	Radio	Newspaper
Elementary school	20	15	15
High school	15	12	3
College	10	8	2

98. On November 18, 1980, the *Wall Street Journal* published a Gallup survey of the opinions of 782 chief executives of US corporations. The 782 chief executives represent samples of 282 from large firms, 300 from medium-sized firms and 200 from small firms. Frank Allen, a staff reporter for the *Wall Street Journal,* used a questionnaire to ask "How many people in your company are capable of doing your job as chief executive?" The results are presented in the table.
Use the chi-square statistic to test whether "number of people capable of doing your job" is independent of "size of firm" at $\alpha = .05$.

A 6 × 3 Contingency table for 782 chief executives' responses

Suitable successors	Large firms	Medium firms	Small firms
1	6 %	10 %	22 %
2	14	27	30
3	24	26	18
4 or 5	30	21	8
6 or more	22	11	4
Don't know	4	5	18

Source: Wall Street Journal, November 18, 1980. Reprinted by permission of the Wall Street Journal, © 1980 Dow Jones & Company, Inc. All Rights Reserved Worldwide

99. Money magazine (Money, March 2003) reports percentage returns and expense ratios for top bond funds under four categories: US government (G), high-yield corporate (H), tax exempt (T), world bond funds (W). Can we conclude that there is significant difference in the mean expense ratio among the four types of bond funds? Do a 5 % test, using a one-way ANOVA.

G	H	T	W
5.0	9.7	5.6	4.5
4.9	8.8	5.1	4.2
4.5	7.6	4.5	7.4
3.6	7.1	3.0	8.8
3.9	7.1	4.5	3.4
4.4	8.0	3.6	4.0
4.5	9.7	5.0	4.4
4.9	8.4	4.2	3.7

100. Use the data from problem **99**. At the $\alpha = 0.05$, use Scheffé's multiple comparison to test for the difference between any pair.
101. The rates of returns data from year 2000 to 2009 in Table 4.15 is listed below:

Year	JNJ	MRK	S&P 500
2000	0.140	0.412	0.075
2001	−0.431	−0.357	−0.163
2002	−0.078	−0.013	−0.168
2003	−0.021	−0.158	−0.029
2004	0.249	−0.272	0.171
2005	−0.032	0.037	0.068
2006	0.122	0.418	0.086
2007	0.035	0.367	0.127
2008	−0.076	−0.451	−0.174
2009	0.108	0.254	−0.223

At the $\alpha = 0.05$, can we conclude that the two stocks and the general market generate about the same average rate of return by using a one-way ANOVA?
102. Use the data in Problem 9. At the $\alpha = 0.05$, do a two-way ANOVA to determine whether the rates of return for the three kinds of stocks/general market are about the same.

Project III: Project for Statistical Inferences Based on Samples

Use the rates of return data presented in Table 12.29 to do the following:

1. Calculate the mean and the standard deviation for JNJ, Merck, and the market.
2. Calculate the confidence intervals for rates of return of JNJ, Merck, and the market.
3. Test whether the average rates of return of JNJ and Merck are significantly different from the market rates of return.
4. Use the data of JNJ, Merck, and the market to perform the ANOVA test and to write an analysis about the results.
5. Test whether market rates of return are normally distributed in accordance with the χ^2 distribution.
6. Download monthly adjusted close price data of JNJ from Yahoo Finance during the period from January 2005 to current month, calculate the rates of return of JNJ, and redo 1–5.

Appendix 1: ANOVA and Statistical Quality Control

In statistical quality control, we can use the ANOVA to measure the system analysis, capability studies control chart set up.[5] In this appendix, we will show how the ANOVA can be used to do measurement system analysis.

Measurement variation can be broken down into two components:

1. Reproducibility variation due to the measurement *system*. It is the variation observed when different operators measure different parts using the same device repeatedly.
2. Repeatability variation due to the measuring *device*. It is the variation observed when the same operator measures the same part with the same device repeatedly.

MINITAB provides a gage R&R (repeatability and reproducibility) study and a gage run chart break for examining measurement variation.

Gage Run Chart

Before running a gage R&R study, you may want to look at your measurement data on a plot. The gage run chart command plots of all the observed measurements for each operator/part combination, letting you *visualize* the repeatability and reproducibility components of the measurement variation.

Gage R&R Study

You can choose between two types of gage R&R studies—the ANOVA method discussed in this chapter and the $\bar{X}$ part and $\bar{R}$ method which have been discussed in Chap. 10:

- The ANOVA method breaks down measurement system variation into reproducibility and repeatability. It also goes one step further and breaks down reproducibility into its operator and operator by part components.

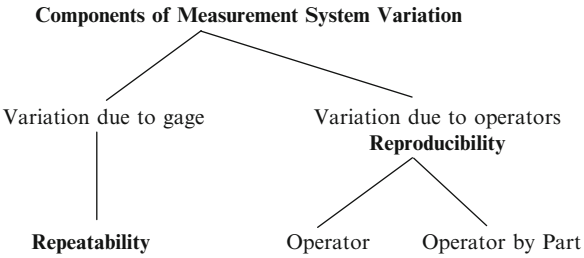
– The operator component is the variation observed between different operators measuring the same set of parts.
– The operator by part component is the interaction between operator and part, that is, the variation among the average part sizes measured by each operator.

[5] The basic information of this appendix is essentially based upon *MINITAB Reference Manual* Release 11 June 1996, pp 10–1 through 10–28.

For example, one operator may get more variation when measuring smaller parts, whereas another operator may get more variation when measuring larger parts.

- The $\bar{X}$ and $\bar{R}$ method also breaks down measurement system variation into reproducibility and repeatability, but not reproducibility into its operator and operator by part components. In this case, reproducibility equals operator.

The ANOVA method provides a more accurate assessment of the measurement system than the $\bar{X}$ and $\bar{R}$ method because it accounts for the operator by part interaction.



Example 12.4 Gage R&R Study: The ANOVA Method. The U-bolts sample measurement data for the auto supply company presented in Table 10.3 of Chap. 10 represent samples from 3 shifts. There are 8 h in each shift. Using the ANOVA method, we can decompose the measurement variation into (1) variation between shifts, (2) variation between hours, (3) variation due to interaction between shift and hour, and (4) errors. The MINITAB output is shown in Fig. 12.6.

If we rearrange the data to represent outputs produced by three operators, each operator produces eight different parts process as shown in Table 12.27

- The variation in terms of the data of Table 12.27 can be decomposed by the ANOVA method into (1) variation between operators, (2) variation between parts, (3) variation due to the interaction between operators and parts, and (4) errors. The MINITAB output of this result is present in Fig. 12.6.

From Fig. 12.6, the F ratios for each source of variation are as follows:

$$\text{Variation between operators } \frac{0.00058}{0.00633} = 0.09$$

$$\text{Variation between parts } \frac{0.01912}{0.00633} = 3.02$$

$$\text{Variation due to interaction } \frac{0.01104}{0.00633} = 1.74$$

Table 12.27 Output measurements with different part number by three different operators

		Operators													
Part number		1				2						3			
1	10.65	10.70	10.65	10.65	10.85	10.75	10.85	10.75	10.85	10.65	10.75	10.80	10.80	10.70	10.75
2	10.60	10.70	10.70	10.75	10.65	10.70	10.75	10.65	10.85	10.80	10.60	10.75	10.75	10.85	1O.70
3	10.60	10.80	10.70	10.75	10.75	10.75	10.80	10.65	10.75	10.70	10.65	10.80	10.85	10.85	10.75
4	10.60	10.70	10.60	10.80	10.65	10.80	10.75	10.90	10.50	10.85	10.85	10.75	10.85	10.65	10.70
5	10.70	10.70	10.75	10.75	10.70	10.65	10.70	10.85	10.75	10.60	10.75	10.80	10.75	10.80	10.65
6	10.90	10.80	10.80	10.75	10.85	10.75	10.70	10.85	10.70	10.80	10.75	10.70	10.60	10.70	10.60
7	10.65	10.65	10.85	10.65	10.70	10.60	10.60	10.65	10.55	10.65	10.50	10.55	10.65	10.80	10.80
8	10.80	10.65	10.75	10.65	10.65	10.65	10.60	10.65	10.60	10.70	10.65	10.70	10.70	10.60	10.65

From the critical values of F in Table A6, it can be concluded that only the variation between parts is significantly different from zero at $\alpha = .05$

MTB > TWOWAY USING DATA IN Cl, A LEVEL IN C2, B LEVEL IN C3

Two-Way Analysis of Variance

Analysis of variance fcr Cl

Source	DF	SS	MS
C2	2	0.00117	0.00058
C3	7	0.13381	0.01912
Interaction	14	0.15450	0.01104
Error	96	0.60800	0.00633
Total	119	0.89748	

MTB > PRINT C-C3

Data Display

Row	C1	C2	C3
1	10.65	1	1
2	10.70	1	1
3	10.65	1	1
4	10.65	1	1
5	10.85	1	1
6	10.60	1	2
7	10.70	1	2
8	10.70	1	2
9	10.75	1	2
10	10.65	1	2
11	10.60	1	3
12	10.80	1	3
13	10.70	1	3
14	10.75	1	3
15	10.75	1	3
16	10.60	1	4
17	10.70	1	4
18	10.60	1	4
19	10.80	1	4
20	10.65	1	4
21	10.70	1	5
22	10.70	1	5
23	10.75	1	5
24	10.75	1	5
25	10.70	1	5
26	10.90	1	6
27	10.80	1	6
28	10.80	1	6
29	10.75	1	6

Fig. 12.6 MINITAB output of two-way analysis

Row	C1	C2	C3
30	10.85	1	6
31	10.65	1	7
32	10.65	1	7
33	10.85	1	7
34	10.65	1	7
35	10.70	1	7
36	10.80	1	8
37	10.65	1	8
38	10.75	1	8
39	10.65	1	8
40	10.65	1	8
41	10.75	2	1
42	10.85	2	1
43	10.75	2	1
44	10.85	2	1
45	10.65	2	1
46	10.70	2	2
47	10.75	2	2
48	10.65	2	2
49	10.85	2	2
50	10.80	2	2
51	10.75	2	3
52	10.80	2	3
53	10.65	2	3
54	10.75	2	3
55	10.70	2	3
56	10.80	2	4
57	10.75	2	4
58	10.90	2	4
59	10.50	2	4
60	10.85	2	4
61	10.65	2	5
62	10.70	2	5
63	10.85	2	5
64	10.75	2	5
65	10.60	2	5
66	10.75	2	6
67	10.70	2	6
68	10.85	2	6
69	10.70	2	6
70	10.80	2	6
71	10.60	2	7
72	10.60	2	7
73	10.65	2	7
74	10.55	2	7
75	10.65	2	7
76	10.65	2	8

Fig. 12.6 (continued)

Row	C1	C2	C3
77	10.60	2	8
78	10.65	2	8
79	10.60	2	8
80	10.70	2	8
81	10.75	3	1
82	10.80	3	1
83	10.80	3	1
84	10.70	3	1
85	10.75	3	1
86	10.60	3	2
87	10.75	3	2
88	10.75	3	2
89	10.85	3	2
90	10.70	3	2
91	10.65	3	3
92	10.80	3	3
93	10.85	3	3
94	10.85	3	3
95	10.75	3	3
96	10.85	3	4
97	10.75	3	4
98	10.85	3	4
99	10.65	3	4
100	10.70	3	4
101	10.75	3	5
102	10.80	3	5
103	10.75	3	5
104	10.80	3	5
105	10.65	3	5
106	10.75	3	6
107	10.70	3	6
108	10.60	3	6
109	10.70	3	6
110	10.60	3	6
111	10.50	3	7
112	10.55	3	7
113	10.65	3	7
114	10.80	3	7
115	10.80	3	7
116	10.65	3	8
117	10.70	3	8
118	10.70	3	8
119	10.60	3	8
120	10.65	3	8

Fig. 12.6 (continued)

Part IV
Regression and Correlation: Relating Two or More Variables

Part III of this book deals with statistical inference based on samples. This part continues the discussion of inferential statistics but focuses on the relationship between two or more variables, using regression and correlation analyses. Regression analysis is one of the analytical tools most frequently used in many areas of business and economics. Chapters 13 and 14 focus on simple regression and correlation analysis. Chapter 15 discusses regression analysis, and Chap. 16 explores the subject further.

Part IV includes applications and examples in accounting, economics, finance, marketing, and other areas of business.

Chapter 13
Simple Linear Regression and the Correlation Coefficient

Chapter Outline

Key Terms

Simple regression analysis	Normal equations
Multiple regression analysis	Sum of squared deviations
Dependent variable	Autocorrelated residuals
Response variable	Best linear unbiased estimator (BLUE)
Independent variable	Standard error of residuals
Explanatory variable	Coefficient of determination
Linear model	Total variation
Regression model	Explained variation
Intercept	Unexplained variation
Slope	Sample standard deviation of error
Regression coefficient	Term
Scatter diagram	Degrees of freedom
Free-hand drawing method	Correlation analysis
Method of least squares	Bivariate normal distribution
Standard deviation of error term	American call option portfolio weight

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_13,

13.1 Introduction

In Sect. 6.9, we used correlation to provide a measure of the strength of any linear relationship between a pair of random variables X and Y. The random variables are treated perfectly symmetrically; that is, "the correlation between X and Y" is equivalent to "the correlation between Y and X." In this chapter, we first discuss the linear relationship between a pair of variables without perfect symmetry. In other words, we assume that Y is a dependent variable and X an independent variable: Y depends on X. Then we discuss the bivariate normal relationship and concepts related to the correlation coefficient.

Regression analysis is perhaps the statistical technique used most frequently to analyze the relationship between two or more variables in business and economics. This technique deals with the way one variable tends to change as one or more other variables change. In this chapter and the next, we will consider a regression relationship in which Y depends on only one variable X. Examples of this relationship include how sales (Y) vary with advertising expenditures (X), how quantity demanded (Y) varies with prices (X), and the relationship between corporate profit (Y) and R&D spending (X). Because all these cases deal with the relationship between two variables only, we call this kind of relationship a *simple regression analysis*. In Chap. 15, we will extend regression analysis to cases where more than two variables come into play, such as the relationship among sales, price, advertising expenditures, and perhaps even growth of gross national product. A regression analysis that involves more than two variables is called a *multiple regression analysis*. In Chap. 16, other important techniques and issues related to simple and multiple regression are discussed in detail.

In this chapter, we first discuss the regression model and population parameters and then distinguish the sample regression model from the population regression model. The least-squares estimation of population parameters, standard assumptions for linear regression, standard error of estimate, and coefficient of determination are investigated. Finally, we explore the bivariate normal distribution and correlation analysis. The relationships among simple regression, slope, and correlation coefficient are also discussed.

13.2 Population Parameters and the Regression Models

To study the relationship between two variables, we must distinguish between the dependent variable, denoted by Y, and the independent variable, denoted by X. Here the term *dependent variable* means that the values of an estimated variable depend on the values of another variable. The dependent variable may also be known as the *response variable*. The *independent variable*, which is also known as the

Table 13.1 Population height and weight data for children

x (inches)	y (pounds)					$E(Y\|X)$
55	91	92	93	94	95	93
56	92	94	95	97	97	95
57	92	95	96	99	103	97
58	94	97	98	100	106	99
59	95	97	100	102	111	101
60	94	99	101	104	117	103

explanatory variable, is used to explain the dependent variable.[1] The value of an explanatory variable normally offers at least a partial explanation of the behavior of the dependent variable.

For example, in economic analysis when we investigate the relationship between income and consumption of goods and services, consumption is the dependent variable (Y) and income the independent variable (X). Consumption (consumer spending) depends on, and is determined by, level of income. Let's use regression analysis to consider the relationship between height and weight in a group of children. This set of common-sense data will be used in both Chaps. 13 and 14 to demonstrate how simple regression analysis can be done intuitively. In Chap. 14, business and economic applications of simple regression are also discussed.

13.2.1 Data Description

Suppose we have a group of children who are classified according to their height, as shown in Table 13.1. The population consists of 30 pairs of observations: (55 in., 91 lb), (55 in., 92 lb),. . ., (60 in., 117 lb). Figure 13.1 is a graph of these observations. Note that these groups are formed according to fixed heights, such as 55 in. and 56 in., and that each group, or subpopulation, has 5 pairs of observations. There are 6 subpopulations corresponding to the fixed variable heights (X). We shall say that we have a collection, or family, of subpopulations. The average value of Y in each subpopulation is called the expected value for a given height X. It is written $E(Y|X)$ and is given in the last column of Table 13.1. For example, the average value of Y for a height of 60 in. is

$$E(Y|X = 60) = \frac{94 + 99 + 101 + 104 + 117}{5} = 103 \text{ lb}$$

Using a similar approach, we can calculate the subpopulation means of all the other groups. They are graphed as the straight line ABC in Fig.13.1.

[1] For instance, the equation $y = x + 3$ is a linear model with x as the independent variable and y as the dependent variable. The variable x is considered independent because it is predetermined. For any given value of x, we can find a corresponding value of y, so the value of y is dependent on the value of x. When x is equal to 4, y is equal to 7. Strictly speaking, the word *independent* implies that the values of this variable are preassigned and that the values of the dependent variable follow, at least in part, from this preassignment.

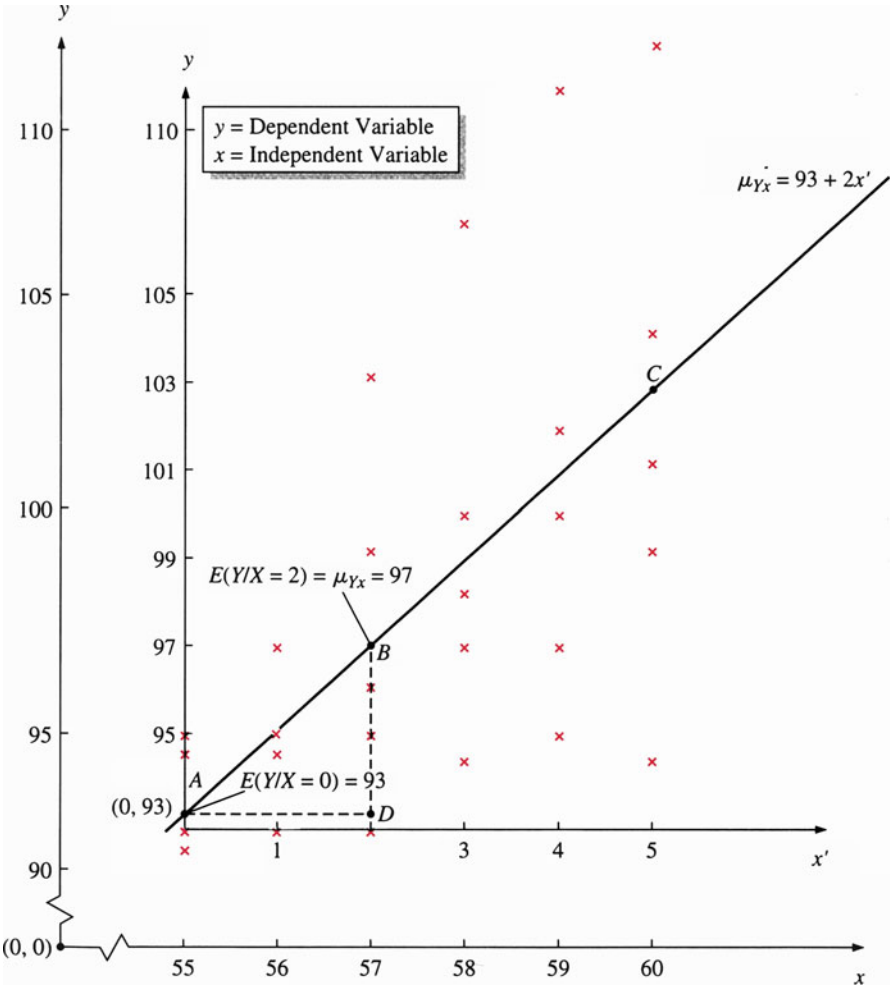


Fig. 13.1 Linear relationship between weights and heights for 30 children

13.2.2 Building the Population Regression Model

Let us now focus our attention on the subpopulation corresponding to $X = 57$ in.

$$E(Y|X = 57) = \frac{92 + 95 + 96 + 99 + 103}{5} = 97$$

The $y = 103$ lb in this subpopulation corresponding to $x = 57$ in. deviates from $E(Y|X)$ by

$$y - E(Y|X = 57) = 103 - 97 = 6 \text{ lb}$$

We will express such deviations as ε, so $y = 103$ lb can be expressed as

$$y = E(Y|X = 57) + \varepsilon$$

where ϵ is the error term, which is a random variable. This is a general expression for individual Y-values of the $X = 57$ subpopulation. That is, when $\varepsilon = -5$,

$$y = E(Y|X = 57) + \varepsilon = 97 - 5 = 92$$

When $\varepsilon = -2$, $y = 95$; when $\epsilon = -1$, $y = 96$; when $\varepsilon = 2$, $y = 99$; and when $\varepsilon = 5$, $y = 102$.

The various y-values in each subpopulation can be expressed in a similar manner.

$$\begin{aligned} y_1 &= E(Y|X_1 = 55) + \varepsilon_1 = 93 + \varepsilon_1 \\ y_2 &= E(Y|X_2 = 56) + \varepsilon_2 = 95 + \varepsilon_2 \\ &\vdots \\ y_{30} &= E(Y|X_{30} = 60) + \varepsilon_{30} = 103 + \varepsilon_{30} \end{aligned}$$

In general the ith value of Y is expressed as

$$Y_i = E(Y_i|X_i = x_i) + \varepsilon_i \tag{13.1}$$

where $E(Y_i|X_i)$ represents the expected value of those Y for which X is equal to the specific value x_i, and ε_i is the error term associated with ith observation in the population regression.

$E(Y_i|X_i = x_i)$ gives us a straight line, as shown in Fig. 13.1, so we can express $E(Y_i|X_i = x_i)$ as

$$E(Y_i|X_i = x_i) = \alpha + \beta x_i \tag{13.2}$$

where α is the y-intercept, β is the slope, and x_i is the ith independent variable. This is called a linear function because the resulting curve is a straight line. Let

$$\begin{aligned} E(Y|X = x) &= \mu_{Yx} \\ &= \alpha + \beta x \end{aligned} \tag{13.3}$$

This equation represents a linear relationship between $E(Y|X = X)$ and x for all data, whereas Eq. 13.2 represents only the relationship for a specific pair of data. In addition, Eq. 13.3 represents the conditional population mean as presented by line ABC in Fig. 13.1.

Table 13.2 Worksheet for calculating μ_{yx}

Heights, x_i (inches)	$x'_i = x_i - 55$	μ_{yx}
55	0	93
56	1	95
57	2	97
58	3	99
59	4	101
60	5	103

Equation 13.3 represents a straight line with slope β and intercept α. Then the slope for line *ABC can* be interpreted as[2]

$$\beta = \frac{97 - 93}{57 - 55} = 2$$

The parameter β, the slope, measures the change in Y resulting from a change in X. It is calculated by dividing the change in Y by the change in X. One way of interpreting a slope of 2 is to say that if the independent variable X is changed by 1 unit, the dependent variable changes by + 2 units. To obtain the y-intercept, we shift the origin from (0, 0) to A (0, 93). In other words, we let the origin $x = 0$ for the height of 55 in.; then $\alpha = 93$. Hence, the straight line used to describe ABC is $\mu_{yx} = 93 + 2x'$. Substituting $x' = 0, 1, 2, 3, 4$, and 5 into μ_{yx}, we obtain the results indicated in Table 13.2. By combining Eqs. 13.1 and 13.2, we can express an individual value of Y as

$$Y_i = \alpha + \beta x_i + \varepsilon_i \tag{13.4}$$

where Y_i and x_i represent the ith value for Y and x, respectively.

Equations 13.1, 13.2, and 13.4 summarize all the data in the population and are called the *linear model* (or *regression model*). Equation 13.3 is called the *regression function*; it shows the relationship between the expected values of Y and the independent values X. The *y-intercept* α and the *slope* β are called *regression coefficients* (parameters).[3]

Using the population data listed in Table 13.2, we have our population regression line (Eq. 13.5) and our population regression model (Eq. 13.6) for describing the relationship between weights and heights:

$$\mu_{Yx} = 93 + 2x'_i \tag{13.5}$$

$$Y_i = 93 + 2x'_i + \varepsilon_i \tag{13.6}$$

[2] From ΔABD, the slope of ABC can be defined as $\beta = BD/AD = (97-93)/(57-55) = 2$.

[3] For an illustration of the meaning of the model, let x be the amount of advertising and Y be the amount of sales. Equation 13.3 tells us that, given a certain amount of advertising, the expected amount of sales is $\mu_{yx} = \alpha + \beta x$.

Population regression, as indicated in Eq. 13.5, represents conditional mean values. In Fig. 13.1 the population mean value for a height of 57 in. is seen to be $\mu_{y,57} = 97$ lb. In other words, the average weight for all children with a height of 57 in. is 97 lb. The value is calculated by substituting $x' = (57 - 55) = 2$ in. into the population regression line as follows: $\mu_{y,57} = 93 + 2(2) = 97$ lb.

In this section we have shown that in a simple regression analysis, two population regression parameters are to be calculated. Our assumption that α and β are known is, of course, an unrealistic one. Usually, α and β can only be estimated in terms of sample data.

13.2.3 Sample Versus Population Regression Model

If we have a large amount of information from a population to analyze, it may not be possible (or desirable) to obtain this specific information on each element in the population. Under these circumstances we generally use a *sample* to estimate the population parameters of the regression line in accordance with n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. In the case of simple regression analysis, two population parameters, α and β, need to be estimated. Once we have estimates of α and β, we can derive an estimate of μ_{yx} for any specified value of X. The sample regression line used to estimate a and β and to predict μ_{yx} can be defined as

$$y = a + bx \tag{13.7}$$

where a and b are the intercept and slope to be estimated in terms of sample data.

Let us explore how this sample regression line is related to the population regression line described by Eq. 13.3. The sample value of a is used to estimate α, and the sample value of b is used to estimate β. The values of a and b, together with a given value of X, yield an estimated value of Y that we can use to estimate the population value μ_{yx} defined in Eq. 13.3.

We can add the subscript i to these variables of Eq. 13.7 to indicate specific values, just as we did for the population regression line. Thus, if x_i is a specific value of X, the equation for estimating α and β is

$$y_i = a + bx_i + e_i \tag{13.8}$$

where a and b are intercept and slope in a sample regression with error term e_i.

Equation 13.8 yields a sample regression line that can be used to estimate parameters of the population regression line defined in Eq. 13.4. As we will see in the next section, we take n pairs of sample observations to estimate α and β.

Table 13.3 Sample height and weight data for children

x_i (inches)	y_i (pounds)
55	92
56	95
57	99
58	97
59	102
60	104

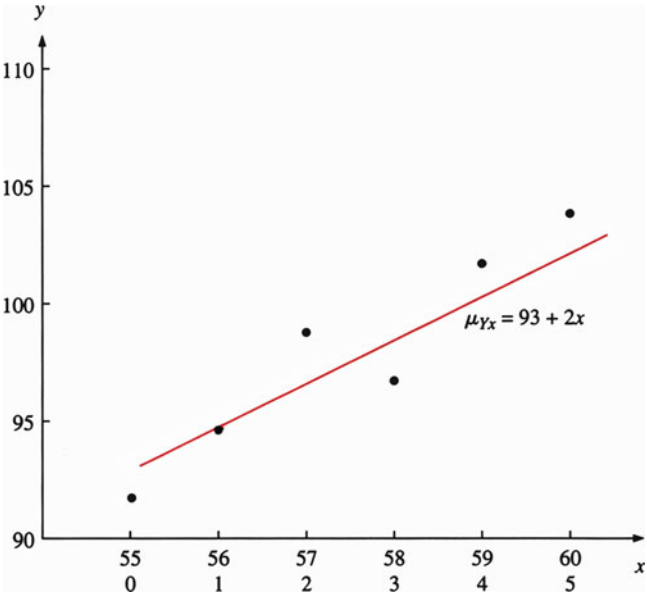


Fig. 13.2 Scatter diagram

13.3 The Least-Squares Estimation of α and β

In this section we discuss the scatter diagrams, method of least squares, and how α and β are estimated.

13.3.1 Scatter Diagram

Using the hypothetical population we first met in Table 13.1, we select a random sample, for simplicity choosing one pair from each subpopulation. The sample is given in Table 13.3. Figure 13.2 is a graph of these observations that is called a *scatter diagram*. We can estimate the model without a diagram, but the scatter diagram gives us a preliminary idea of the shape of the regression function. For these six observations, we observe from the scatter diagram that the relationship

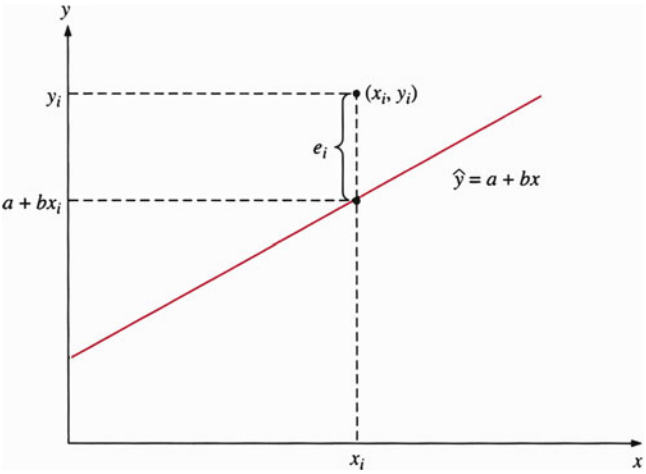


Fig. 13.3 Measurement of ith residual term, $e_i = y_i - (a + bx_i)$

is linear. In addition, the scatter diagram enables us to make rough estimates of α and β.

We would like to use a line to show the relationship between x and y in Fig. 13.2. A simple method of drawing a line to describe the relationship between x and y is the so-called *free-hand drawing method*, whereby we just draw a line in accordance with our best judgment about the relationship between x and y. However, the free-hand drawing method does not necessarily give systematic and objective estimates for α and β. Furthermore, the free-hand method provides no way of measuring sampling errors, which are always important in forming confidence intervals or doing tests of hypotheses on population parameters. From Eq. 13.4,

$$\varepsilon_i = Y_i - \alpha - \beta x_i$$

where ε_i represents the error term for the ith observation in population regression.

In a similar manner, we can define the sample residual (error) term as

$$e_i = y_i - a - bx_i \tag{13.9}$$

where e_i is used to measure the distance from the point (x_i, y_i) to the line, as indicated in Fig. 13.3.

What we need now is a mathematical procedure for determining the sample regression line that best fits the data. The most reasonable approach is to find the values of a and b such that the estimated values of dependent variable $\hat{y}$ (in the equation $\hat{y} = a + bx$) are as close as possible to the observed values y.

13.3.2 The Method of Least Squares

A method of minimizing *the sum of squared deviations* is used as a criterion for finding values of a and b. The smaller e_i is, the closer $\hat{y}$ is to the actual y_i-value. Put another way, the smaller e_i is, the better the fit of the regression line is.

Because a small value for e_i is desirable, we wish to find values for a and b that will make e_i as small as possible. In other words, we find the line of best fit in regression analysis by determining the values of a and b that minimize *the sum of the squared residuals*. This procedure is known as the *method of least squares*. It is accomplished as follows:

$$\text{Minimize} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13.10}$$

The sample regression line determined by minimizing $\sum_{i=1}^{n} e_i^2$ is called the *least-squares regression line*. Because $\hat{y}_i = a + bx_i$, minimizing

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

is equivalent to minimizing

$$\sum_{i=1}^{n} e_i{}^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{13.11}$$

That is, we find a and b such that the sum of squared deviations $\sum_{i=1}^{n} e_i^2$, taken over the sample values, is at a minimum. We estimate a and b by the two *normal equations*. (The derivation of the normal equations is shown in Appendix 1.)

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \tag{13.12}$$

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \tag{13.13}$$

Equations 13.12 and 13.13 can be regarded as a two-equation simultaneous equation system. The two unknowns are the estimates a and b (not y and x) because we must choose a and b from among an infinite possible set of values, given the sample of values of y_i and x_i.

13.3.3 Estimation of Intercept and Slope

To estimate the intercept, we divide Eq. 13.12 by n and rearrange terms.

$$a = \bar{y} - b\bar{x} = \frac{\sum_{i=1}^{n} y_i - b\left(\sum_{i=1}^{n} x_i\right)}{n} \tag{13.14}$$

Equation 13.14 implies that the intercept of a simple regression is the mean of y ($\bar{y}$) minus the slope (b) times the mean of x ($\bar{x}$). Here b is yet to be estimated.

To estimate the slope b, we substitute Eq. 13.14 into 13.13, and, letting $\Sigma_{i=1}^{n} x_i = n\bar{x}$, we obtain

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)} \tag{13.15}$$

Although the formulas given in Eqs. 13.14 and 13.15 are useful, in a practical sense it is just as easy to use Eqs. 13.12 and 13.13 directly. These two equations require only the solution of two equations (linear) in two unknowns.

Alternatively, we replace x_i by its deviation from $(x_i - \bar{x})$ in Eq. 13.13 and obtain

$$\sum_{i=1}^{n} (x_i - \bar{x}) y_i = a\sum_{i=1}^{n} (x_i - \bar{x}) + b\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Because the first term on the right-hand side of this equation is zero, the equation immediately implies that[4]

$$\begin{aligned} b &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \\ &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})/n \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2/n \end{aligned} \tag{13.16}$$

[4] The second equality of Eq. 13.16 holds because

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n} (x_i - \bar{x}) y_i - \bar{y}\sum_{i=1}^{n} (x_i - \bar{x}) \\ &= \sum_{i=1}^{n} (x_i - \bar{x}) y_i \end{aligned}$$

Table 13.4 Procedure for calculating a and b

	(1)	(2)	(3)	(4)	(5)	(6)
	x_i (inches)	y_i (pounds)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
	55	92	−2.5	−6.1667	15.4168	6.25
	56	95	−1.5	−3.1667	4.7501	2.25
	57	99	–.5	.8333	–.4167	.25
	58	97	.5	−1.1667	–.5834	.25
	59	102	1.5	3.8333	5.7499	2.25
	60	*104*	*2.5*	*5.8333*	*14.5833*	*6.25*
Sum	345	589	0	0	39.5	17.50
Mean	57.5	98.1667				

where s_{xy} represents the sample covariance between x and y (as discussed in Sect. 6.9 of Chap. 6) and s_x^2 represents the sample variance of x.

Example 13.1 Relationship Between Height and Weight. Using the sample data of Table 13.3, we will illustrate how Eqs. 13.14 and 13.16 can be used to estimate the least-squares regression line and its parameters.[5] Columns (1) and (2) of Table 13.4 give the hypothetical data of heights and weights for 6 children.

The sums in columns (5) and (6) of Table 13.4 give us the information we need to calculate b via Eq. 13.16.

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{39.5}{17.50} = 2.2571$$

This implies that each 1-in. increase in height spells a 2.2571-lb increase in weight.

Using this value of b and the means of x and y shown in Table 13.4, we obtain the following value of α:

$$a = \bar{y} - b\bar{x} = 98.1667 - (2.2571)(57.5) = -31.6166$$

Hence, the least-squares regression line for this example is

$$\hat{y}_i = -31.6166 + 2.2571x_i \tag{13.17}$$

[5] In general, a sample of 6 would not be sufficient. We use a small sample here for computational ease only.

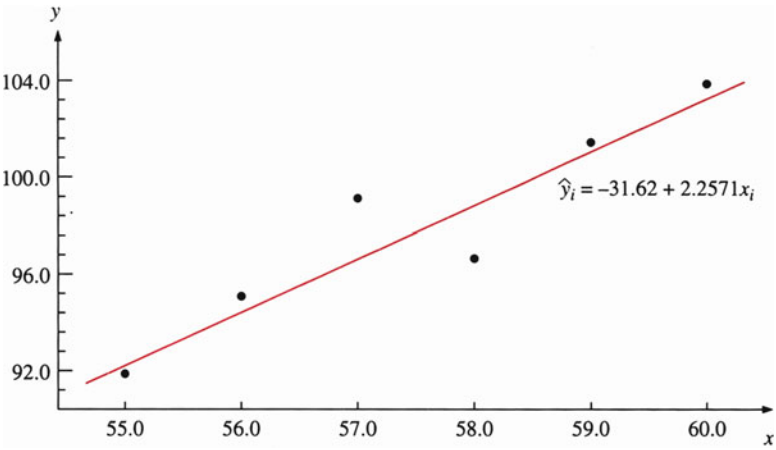


Fig. 13.4 Scatter diagram and regression line

Table 13.5 Observations of y_i, $\hat{y}_i$, and e_i

y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
92	92.5239	–.5239
95	94.7810	.2190
97	99.2952	−2.2952
99	97.0381	1.9619
102	101.5523	.4477
104	103.8094	.1906

where $\hat{y_i}$ represents the estimated regression line, as indicated in Fig. 13.4. Substituting $x_i = 55, 56, 57, 58, 59$, and 60 into Eq. 13.17, we obtain 6 estimated values of $\hat{y_i}(y_i's)$.

$$\begin{aligned}
\hat{y}_1 &= -31.6166 + (2.2571)(55) = 92.5239 \\
\hat{y}_2 &= -31.6166 + (2.2571)(56) = 94.7810 \\
\hat{y}_3 &= -31.6166 + (2.2571)(57) = 97.0381 \\
\hat{y}_4 &= -31.6166 + (2.2571)(58) = 99.2952 \\
\hat{y}_5 &= -31.6166 + (2.2571)(59) = 101.5523 \\
\hat{y}_6 &= -31.6166 + (2.2571)(60) = 103.8094
\end{aligned}$$

The regression line of Fig. 13.4 was determined by the method of least squares, so there is no other line that could be drawn such that the sum of squared residuals between the points and the line (measured in a vertical direction) would be smaller than this line. The residuals and the estimated values of y_t for the 6 sample points are summarized in Table 13.5.

Example 13.2 The Relationship Between the Price of Gasoline and the Price of Crude Oil. The Organization of the Petroleum Exporting Countries (OPEC) has tried to control the price of crude oil since 1973. From the mid-1980s to 1990s, the price of a barrel of crude oil was generally under $25/barrel. However, the price of crude oil rose dramatically in the 2000s. In 2008, the oil price reached the record high price. As a result, motorists were confronted with a similar upward spiral of gasoline prices. The following table presents a gallon of regular leaded gasoline and a barrel of crude oil in terms of the average value at the point of production during 1990–2010 (data from *U.S. Energy Information Administration*, EIA).

Price of gasoline and crude oil

Year, i	Gasoline y ($/gallon)	Crude oil x ($/barrel)
1990	1.299	24.53
1991	1.098	21.54
1992	1.087	20.58
1993	1.067	18.43
1994	1.075	17.2
1995	1.111	18.43
1996	1.224	22.12
1997	1.199	20.61
1998	1.03	14.42
1999	1.136	19.34
2000	1.484	30.38
2001	1.42	25.98
2002	1.345	26.18
2003	1.561	31.08
2004	1.852	41.51
2005	2.27	56.64
2006	2.572	66.05
2007	2.796	72.34
2008	3.246	99.67
2009	2.353	61.95
2010	2.782	79.48

Source: http://www.eia.gov/

To investigate the relationship between the price of a gallon of gasoline and the price of a barrel of crude oil, we estimate the following regressive line:

$$y_i = a + bx_i \tag{13.18}$$

where y_i and x_i represent a gallon of gasoline and a barrel of crude oil in ith year, respectively.

Based on the data listed in the table, we first obtain

$$\sum_{i=1}^{20} y_i = 35.032, \sum_{i=1}^{20} x_i = 788.46, \sum_{i=1}^{20} {y_i}^2 = 68.16171,$$
$$\sum_{i=1}^{20} {x_i}^2 = 41,825.53, \text{and } \sum_{i=1}^{20} x_i y_i = 1,658.53.$$

Substituting this information into Eqs. 13.15 and 13.14, we obtain slope and intercept estimates as

$$\begin{aligned} b &= \frac{20(1,658.53) - (788.46)(35.032)}{20(41,825.53) - (788.46)^2} \\ &= 0.025829 \\ a &= \frac{35.032}{20} - \frac{(0.025829)(788.46)}{20} \\ &= 0.7334 \end{aligned}$$

Substituting estimated a and b into Eq. 13.7, we obtain the estimated regression line as $\hat{y}_i = 0.7334 + 0.02583x_i$.

13.4 Standard Assumptions for Linear Regression

To obtain some desirable properties for the estimators of a regression relationship, we often make five standard assumptions for the standard population regression $Y_i = \alpha + \beta x_i + \epsilon_i$. We shall discuss first the assumptions and then their implications.

Assumption A. Either x_i are fixed numbers (set, e.g., by the experimenter) or they are random variables that are statistically independent of the random variable ϵ_i whose values have been observed (random).

Assumption B. The random variable ϵ_i is assumed to be normally distributed.

Assumption *C*. The random variable ϵ_i is assumed to have a mean of zero; that is, $E(\epsilon_i) = 0$ for $i = 1.2,\ldots, n$. This assumption implies that the mean value of Y given X, $E(Y \mid X)$, is $E(Y \mid X = x_i) = \alpha + \beta x_i$. This assumption implies that there are no omitted variables associated with the population regression specification. (The issue of specification error associated with regression analysis will be discussed in Chap. 16.)

Assumption D. The random variables ϵ_i are assumed to be statistically independent of one another so that $E(\epsilon_i\ \epsilon_j) = 0$ for $i \neq j$. This assumption implies that no correlation exists among errors. (If the errors are correlated over time for time-series data, then we call these kinds of errors autocorrelated errors. This issue will be discussed in Chap. 16.)

Assumption E. The random variables ϵ_i all have constant variance, say σ_ϵ^2, so $E(\varepsilon_i^2) = \sigma_\epsilon^2$, for $i = 1,2,\ldots, n$. In other words, the population error variance is constant over all values of x_i.

Now let's consider the implications of these five assumptions. Assumption *A* holds if x consists of a fixed number because the covariance of a random variable and a constant is always zero. In addition, it should be noted that the constant has no variation from its fixed value. When x_i is a random variable, this assumption may be violated. If x, cannot be measured precisely, then there exists an error-invariable problem; x_i and ϵ_i are not independent of one another.[6]

Assumptions *B* through *E* concern the error term (ϵ_i) in the regression equation. Assumption *B* assumes that the difference between Y_i and their conditional expectations ($\alpha + \beta x_i$) is normally distributed. This assumption is needed only when statistical tests of significance are conducted. Assumption *C* means that for a given x_i the difference between Y_i and its conditional mean ($\alpha + \beta x_i$) is sometimes positive and sometimes negative but on average is zero.

Assumption *D* means that the error of one point in the population cannot be related systematically to the error of any other point in the population. In other words, knowledge about the magnitude and sign of one or more errors does not help us predict the magnitude and sign of any other error. This assumption is frequently violated in time-series analysis, which is discussed in detail in Chap. 18.

Finally, Assumption *E* means that the random errors all have the same variance. Figure 13.5 shows what the error terms should look like with a constant variance and with one that varies.

In summary, assumptions *B* through *E* imply that the random variable ϵ_i is normally, identically, and independently distributed with mean zero and variance σ_ϵ^2. If all the assumptions are true, then the estimators of α and β as determined by the least-squares method are *best linear unbiased estimators (BLUE)*. Essentially, BLUE means that the estimates of the parameters are best because the error variances of least-squares estimators are smaller than those of any other unbiased estimators. *Linear* means that the estimators are a linear function of the observed values of the dependent variable Y. The estimators are said to be unbiased because the expected value of each sample coefficient is equal to the population parameter.

The implications of assumptions *B*, *C*, and *E* are apparent in Fig. 13.6, which shows distributions of errors for three particular values of x: x_1, x_2, and x_3. Note that the relative frequency distributions of errors are normally distributed with mean zero and constant variance σ_ϵ^2. The straight line, shown in Fig. 13.6, plots $E(Y_i \mid X_i = x_i)$ as

$$E(Y_i|X_i = x_i) = \alpha + \beta x_i$$

[6] For instance, if in economic or business research, current instead of permanent income is used as the independent variable in estimating consumption function, then there are proxy errors associated with income measurements, as discussed in Appendix 14A. If the regression equation is part of interdependent equations, then x_i and ϵ_i also are not independent of each other. However, we will take Assumption A as given.

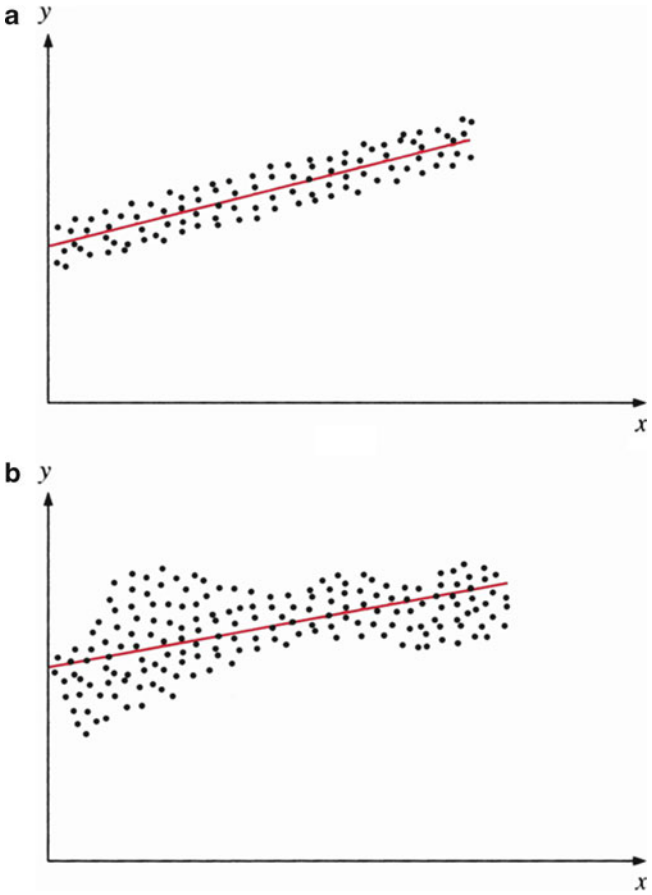


Fig. 13.5 (**a**) Constant and (**b**) nonconstant residual variance

13.5 The Standard Error of Estimate and the Coefficient of Determination

Two alternative measures can be used to measure the goodness of fit for a regression. The *standard error of residuals* is a measure of the absolute fit of the sample points of the sample regression line. The *coefficient of determination* is an index of the relative goodness of fit of a sample regression line. To discuss these two goodness of fit measures, we first need to present some of the components for measuring the variability of y_i in regression analysis.

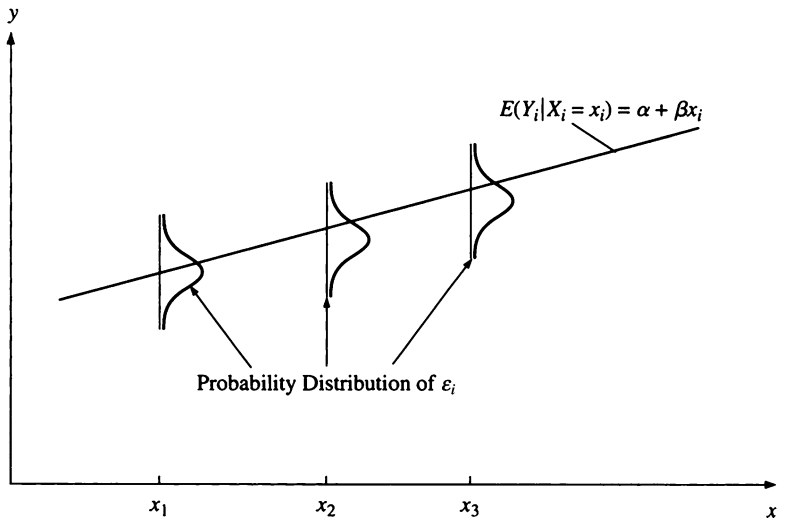


Fig. 13.6 The regression line and the probability distribution of ε_i

13.5.1 Variance Decomposition

In regression analysis, two means are associated with the dependent variable y:

Overall mean $(\bar{y})$
Conditional mean $(\hat{y}_i = a + bx_i)$

Based on these two different means, we can break down the total deviation $(y_i - \bar{y})$ into unexplained deviation $(y_i - \hat{y}_i)$ and explained deviation $y_i - \bar{y}$ as

$$\begin{array}{ccccc} y_i - \bar{y} & = & (y_i - \hat{y}_i) & + & (\hat{y}_i - \bar{y}) \\ \text{Total} & & \text{Unexplained} & & \text{Explained} \\ \text{Deviation} & & \text{Deviation} & & \text{Deviation} \end{array} \tag{13.19}$$

Equation 13.19 implies that the deviation of y_i from its overall mean $(\bar{y})$ can be dissected into two components, $(y_i - \hat{y}_i)$ and $(\hat{y}_i - \bar{y})$. The deviation $(y_i - \hat{y}_i)$ cannot be explained (or accounted for) by the regression line because when x_i changes, both y_i and $\hat{y}_i$ change; hence, it is called the unexplained deviation. However, the deviation $(\hat{y}_i - \bar{y})$ can be explained by the regression line because when x_i changes, $\bar{y}$ remains constant; thus, it is called the explained deviation. The relationship is illustrated in Fig. 13.7. By squaring each deviation and summing overall observations of Eq. 13.19, it can be shown (see Appendix 2) that

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{13.20}$$

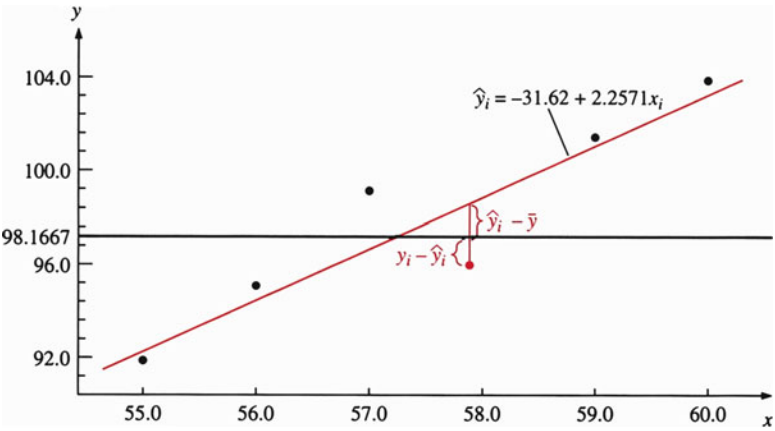


Fig. 13.7 Estimated regression line

This equation implies that the *total variation* of the dependent variable y_i can be dissected into *unexplained variation* and *explained variation*. Alternative terms used to describe these components follow.

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{total variation, sum of squares total (SST)}$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{unexplained variation, sum of squares error (SSE)}$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \text{explained variation, sum of squares due to regression (SSR)}$$

In summary, we have

$$\begin{array}{ccccc} \text{SST} & = & \text{SSE} & + & \text{SSR} \\ \text{Total} & & \text{Unexplained} & & \text{Explained} \\ \text{Variation} & & \text{Variation} & & \text{Variation} \end{array} \tag{13.21}$$

On the basis of Eqs. 13.20 and 13.21, we can define and discuss both the standard error of residuals and the coefficient of determination. We now use our height–weight example to calculate SST and SSE, as shown in Table 13.6. (Note that Table 13.6 is an ANOVA table.)

If we divide both sides of Eq. 13.20 by $(n - 1)$, we have

$$\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-1} + \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n-1} \tag{13.22}$$

Table 13.6 Analysis of variance

(1) Actual y_i	(2) Estimate $\hat{y}_i$	(3) $(y_i - \hat{y}_i)^2 = e_i^2$	(4) $(y_i - \bar{y})^2$
92	92.5239	.27447	38.0282
95	94.7810	.04796	10.0280
99	97.0381	3.84905	.6944
97	99.2952	5.26794	1.3612
102	101.5523	.20043	14.6942
104	103.8094	.03633	34.0274
Total		SSE = 9.67618	SST = 98.8334

Sources of variation	Sum of squares	Degrees of freedom	Mean square
Due to regression	$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k	SSR/k
Residuals	$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	n–k–1	SSE/(n–k–1)
Total	$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$	n–1	SST/(n–1)
Due to regression	89.1572	1	89.1572
Residuals	9.6762	4	2.4191
Total	98.8334	5	19.7667

It can be shown that Eq. 13.22 can be rewritten as[7]

$$\underbrace{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}_{\text{Total variance}} = \underbrace{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-1}}_{\text{residual variance}} + \underbrace{\frac{b^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}_{+\ (\text{slope})^2(\text{variance of independent variable})} \tag{13.23}$$

The residual variance defined in Eq. 13.23 is not an unbiased estimate of the population residual variance, which is discussed in the next section.

In Table 13.6, k represents the number of independent variables. In simple regression analysis, k is equal to 1. In the upper portion of Table 13.6, columns (1) and (2) represent the actual and estimated values listed in Table 13.5. Column (3) represents the squared residuals, and column (4) represents the square of actual observations deviated from the overall mean $\bar{y}$. The lower portion of Table 13.6 represents the results of dissecting the variation, a technique used to calculate the standard error of residuals (estimates) and the coefficient of determination.

[7] Because

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} [a + bx_i - (a + b\bar{x})]^2 = b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

13.5.2 Standard Error of Residuals (Estimate)

The first measure of goodness of fit in regression analysis is called the *sample standard deviation of error term* (s_e).

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} \tag{13.24}$$

where $\text{SSE} = \Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2$. Here s_e is a sample statistic about the goodness of fit of the sample regression line, and s_e^2 represents an unbiased estimate of the variance of the error terms $\left(\sigma_\varepsilon^2\right)$ about the population regression line. From Chap. 9 we know that an unbiased sample variance is calculated by dividing the sum of squared deviations by the degrees of freedom, n–2.

Note that the number of elements that can be chosen freely is called the *degrees of freedom*. In this case there are two sample statistics (a and b) that we must calculate before we can compute the value of $\hat{y}$ (because $\hat{y} = a + bx$). Therefore, only (n–2) observations are *free* to vary if a and b are held constant.

From Table 13.6, s_e for our familiar example involving student height and weight is calculated as follows:

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{9.6762}{4}} = 1.5553$$

The value of s_e can be used to describe the distribution of $\hat{y}_i$ in a manner similar to that which we used in the standard deviation of y (i.e., s_y) to describe the distribution of y. In addition, s_e can be used to describe the distributions of a and b. All these concepts and their applications will be discussed in the next chapter.

13.5.3 The Coefficient of Determination

Alternatively, we can use either Eq. 13.20 or Eq. 13.21 to calculate a relative measure of goodness of fit. If we divide both sides of Eq. 13.21 by SST, we obtain

$$\frac{\text{SST}}{\text{SST}} = 1.0 = \frac{\text{SSE}}{\text{SST}} + \frac{\text{SSR}}{\text{SST}}$$

Using this equation, we can derive the coefficient of determination (R^2) as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{13.25}$$

Because SSE is the unexplained variation in y, the ratio SSE/SST is the proportion of the total variation of the dependent variable y_i that *cannot* be explained by the

regression relation. Similarly, the ratio SSR/SST is the proportion of the total variation that *can* be explained by the regression line. Equation 13.25 is used to explain the relationship between SSR/SST and SSE/SST. In summary, R^2 is used to measure the explanatory power of the independent variable x. In our height and weight example,

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{89.1572}{98.8334} = .9021$$

The R^2 indicated in Eq. 13.25 does not adjust for the degrees of freedom. We have already seen (Table 13.6) that in order to obtain the unbiased s_y^2 and s_e^2, we must divide SST and SSR by the degrees of freedom (n–1) and (n–k–1), respectively. Using these concepts, we can define the *adjusted* coefficient of determination $\bar{R}^2$ as

$$\bar{R}^2 = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} \tag{13.26}$$

For our example, $n = 6$, $k = 1$, SST = 98.8334, and SSE = 9.6762. The adjusted coefficient of determination is

$$\bar{R}^2 = 1 - \frac{9.6762/4}{98.8334/5} = 1 - \frac{2.4191}{19.7667}$$
$$= .8776$$

The magnitude of $\bar{R}^2$ is always less than the magnitude of R^2 because $\bar{R}^2$ has been adjusted for the degrees of freedom.

A MINITAB solution using the height and weight data given in Table 13.4 is shown in Fig. 13.8. This output contains nearly all the calculations performed so far. In particular,

$b = 2.2571$	$\text{SSE}/(n-k-1) = 2.419$
$a = -31.62$	$s_e = 1.555$
$\text{SSR} = 89.157$	$\text{R}^2 = .902$
$\text{SSE} = 9.676$	$\overline{\text{R}}^2 = .878$
$\text{SSR}/k = 89.157$	

13.6 The Bivariate Normal Distribution and Correlation Analysis

In *correlation analysis*, we assume a population where both X and Y vary jointly. Correlation analysis doesn't imply causality as regression analysis does.[8] If both X and Y are normally distributed, then we shall call this joint distribution a *bivariate*

[8] Strictly speaking, regression implies causality only under some *prediction* cases.

```
MTB > READ C1 C2
DATA> 92 55
DATA> 95 56
DATA> 99 57
DATA> 97 58
DATA> 102 59
DATA> 104 60
DATA> END
      6 rows read.
MTB > REGRESS C1 1 C2;
SUBC> DW.

Regression Analysis


The regression equation is
C1 = - 31.6 + 2.26 C2

Predictor       Coef        StDev           T        P
Constant      -31.62        21.39       -1.48    0.213
C2            2.2571       0.3718        6.07    0.004

S = 1.555       R-Sq = 90.2%     R-Sq(adj) = 87.8%

Analysis of Variance

Source       DF          SS          MS          F        P
Regression    1      89.157      89.157      36.86    0.004
Error         4       9.676       2.419
Total         5      98.833


Durbin-Watson statistic = 3.03
```

Fig. 13.8 MINITAB output of Table 13.4

normal distribution.[9] In Chap. 6 we discussed the relationship between two variables in terms of covariance – for example, Cov(*X*,*Y*). Now we will explore correlation analysis.

Both the covariance and the correlation coefficient are designed to measure the degree of a linear relationship between a pair of variables. The covariance is an absolute measure and the correlation coefficient a relative measure in determining the relationship between two variables. The population relationship between two variables can be defined as

$$\sigma_{XY} = \text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \text{ and} \tag{13.27}$$

$$\rho = \sigma_{XY}/\sigma_X\sigma_Y \tag{13.28}$$

[9] The bivariate normal density function will be discussed in Appendix 3.

where μ_x and μ_Y are population means of X and Y, respectively, and σ_x and σ_y are population standard deviations of X and Y, respectively. Cov(X,Y) was discussed in Chap. 6. Equation 13.28 represents the population correlation coefficient ρ, which is standardized by dividing Cov(X,Y) by the product of the population standard deviation of X (that is, σ_x) and the population standard deviation of Y (i.e., σ_y).

Three values of the correlation coefficient ρ that can serve as benchmarks for interpreting a correlation coefficient are $\rho = 1$, $\rho = -1$, and $\rho = 0$. $\rho = 1$ means that two variables X and Y exist in a perfect positive linear relationship; $\rho = -1$ means that two variables X and Y exist in a perfect negative linear relationship; and $\rho = 0$ means that X and Y are not linearly related – that is, they are independent random variables. The association between two variables increases as the magnitude of the correlation coefficient approaches 1. If the absolute value of ρ is less than 1, then the larger (in absolute value) the correlation, the stronger the linear association between two random variables.

13.6.1 The Sample Correlation Coefficient

Sample data of random variables X and Y are used to estimate the population correlation coefficient ρ. The sample statistic associated with ρ is the sample correlation coefficient; it is denoted by the letter r.

$$r = \frac{s_{xy}}{s_x s_y} \tag{13.29}$$

where

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \text{sample covariance}$$

$$s_x = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2} = \text{sample standard deviation of } x$$

$$s_y = \left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2} = \text{sample standard deviation of } y$$

To illustrate the procedure for calculating the sample correlation coefficient, we consider once again the data for our standard height and weight example. Following Tables 13.4 and 13.6, we obtain

$$s_{xy} = (39.5/5) = 7.9$$

$$s_x = \left[\frac{(17.5)}{5}\right]^{1/2} = 1.8708$$

$$s_y = \left[\frac{(98.8334)}{5}\right]^{1/2} = 4.4460$$

Substituting these numbers into Eq. 13.29 yields

$$r = \frac{7.9}{(1.8708)(4.4460)} = .9498$$

13.6.2 The Relationship Between r and b

We can explore the relationship between the value of r and the value of the slope b by comparing Eqs. 13.16 and 13.29, which are reproduced here.

$$b = s_{xy}/s_x^2 \tag{13.30}$$

$$r = s_{xy}/s_x s_y \tag{13.31}$$

Equation 13.29 can be rewritten as

$$r = [(s_{xy})/s_x^2][(s_x)/s_y] = b(s_x/s_y) \tag{13.31a}$$

Because both s_x and s_y are always positive, the sign of r is identical to the sign of b. In other words, a positive correlation must correspond to a regression line with positive slope, and a negative r must correspond to a negative slope.

The magnitude of r is determined by the magnitudes of both b and s_x/s_y. In other words, $b = 1$ does not necessarily imply that $r = 1$, unless $s_x/s_y = 1$. Similarly, $r = 1$ does not necessarily imply that $b = 1$, unless $s_y/s_x = 1$.

13.6.3 The Relationship Between r and R^2

The correlation coefficient (r) is used to measure the relationship between x and y, and the coefficient of determination is used to measure the percentage of the variation of y that is attributable to the variation of x. Hence, it is useful to investigate the relationship between the correlation coefficient r and the coefficient of determination R^2. Squaring both sides of Eq. 13.31a, we have

$$\text{Coefficient of determination } R^2 = \frac{\text{SSR}}{\text{SST}} \tag{13.32}$$

where

$$\text{SST} = \sum_{n=1}^{n} (y_i - \bar{y})^2$$

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

It has been shown (footnote 7) that

$$\text{SSR} = b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{13.33}$$

Substituting Eq. 13.33 and the definition of SST into Eq. 13.25, we obtain

$$\begin{aligned} R^2 &= b^2 \sum_{i-1}^{n} (x_i - \bar{x})^2 \Big/ \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ &= b^2 s_x^2 / s_y^2 \\ &= r^2 \end{aligned} \tag{13.34}$$

From Eq. 13.34 we can conclude that $R^2 = r^2$. In our example, $R^2 = .9021$ and $r^2 = (.9498)^2 = .9021$.

From Eq. 13.33 and the definitions of SST, SSE, and s_e, we can rewrite Eq. 13.24 as

$$s_e = \sqrt{\text{SSE}/(n-2)} \tag{13.35}$$

where

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 - b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \sum_{i=1}^{n} y_i^2 - n\left(\sum_{i=1}^{n} y_i/n\right)^2 - b^2 \left[\sum_{i=1}^{n} x_i^2 - n\left(\sum_{i=1}^{n} x_i/n\right)^2\right] \end{aligned}$$

Example 13.3 The Effect of R&D Spending on a Company's Value. Wallin and Gilman (1986) use a simple linear regression analysis to investigate the effect of research and development (R&D) spending on a company's value.[10] Data for the 20 largest R&D spenders in terms of the 1981–1982 averages are presented in Table 13.7. In this table, y and x represent the price/earnings (P/E) ratio and R&D expenditures/sales (R/S) ratio, respectively. Figure 13.9 illustrates the MINITAB simple linear regression output in terms of the data in Table 13.7.

[10] C. C. Wallin and J. J. Gilman (1986). "Determining the Optimum Level for R&D Spending," *Research Management*, Vol. 14, No. 5, Sept./Oct., 19–24.

Table 13.7 P/E ratio and R/S ratio for top 20 R&D spenders (based on the 1981–1982 average)

Company	P/E ratio, y	R/S ratio, x
1	5.6	.003
2	7.2	.004
3	8.1	.009
4	9.9	.021
5	6.0	.023
6	8.2	.030
7	6.3	.035
8	10.0	.037
9	8.5	.044
10	13.2	.051
11	8.4	.058
12	11.1	.058
13	11.1	.067
14	13.2	.080
15	13.4	.080
16	11.5	.083
17	9.8	.091
18	16.1	.092
19	7.0	.064
20	5.9	.028

Source: Wallin, C.C., Gilman, J.J.: Determining the Optimum Level for R&D Spending. Res. Manage. **14**(5), 19–24 (1986) (adapted from Figure 1, p. 20)

```
MTB > READ C1 C2
DATA> 5.6 .003
DATA> 7.2 .004
DATA> 8.1 .009
DATA> 9.9 .021
DATA> 6.0 .023
DATA> 8.2 .030
DATA> 6.3 .035
DATA> 10.0 .037
DATA> 8.5 .044
DATA> 13.2 .051
DATA> 8.4 .058
DATA> 11.1 .058
DATA> 11.1 .067
DATA> 13.2 .080
DATA> 13.4 .080
DATA> 11.5 .083
DATA> 9.8 .091
DATA> 16.1 .092
DATA> 7.0 .064
DATA> 5.9 .028
DATA> END
     20 rows read.
MTB > CORRELATION C1 C2
```

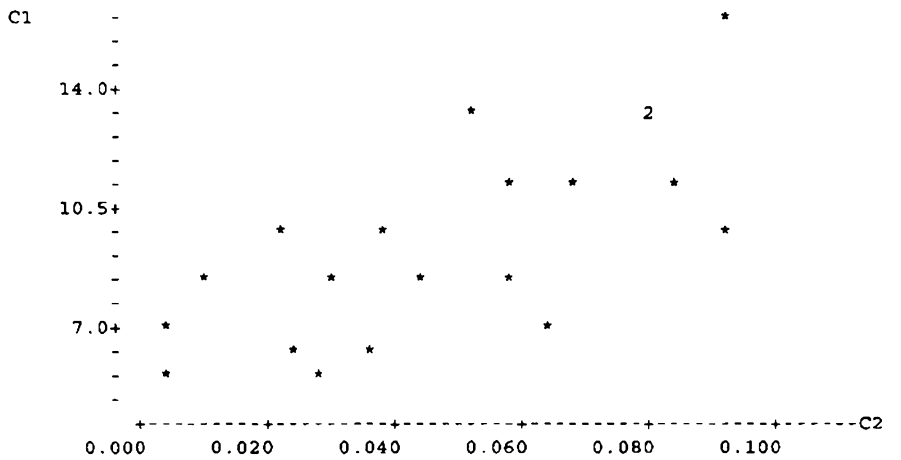
Fig. 13.9 MINITAB output of regression $y(C_1)$on $x(C_2)$

Correlations (Pearson)

```
Correlation of C1 and C2 = 0.726

MTB > GSTD
* NOTE  * Standard Graphics are enabled.
         Professional Graphics are disabled.
         Use the GPRO command to enable Professional Graphics.
MTB > PLOT C1 C2
```

Character Plot




```
MTB > BRIEF 1
MTB > REGRESS C1 1 C2;
SUBC> DW.
```

Regression Analysis

```
The regression equation is
C1 = 5.98 + 74.1 C2

Predictor       Coef       StDev          T        P
Constant      5.9772      0.9174       6.52    0.000
C2             74.07       16.52       4.48    0.000

S = 2.074       R-Sq = 52.7%     R-Sq(adj) = 50.1%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    1      86.404      86.404     20.09    0.000
Error        18      77.414       4.301
Total        19     163.818

Durbin-Watson statistic = 2.58
```

Fig. 13.9 (continued)

Figure 13.9 can be divided into three parts. First, in the data input part, C_1 and C_2 represent y and x, respectively. Second, the output part includes (1) the correlation coefficient between C_1 and $C_2 = .726$, (2) a scatter diagram of plotting C_1 and C_2, and (3) regressing C_1 against C_2.

From the estimate that $r = .726$ and the pattern of the scatter diagram, we can conclude that the P/E ratio is correlated highly with the R/S ratio. The estimated regression line can be denned as

$$\bar{y} = 5.98 + 74.1x$$

In sum, the MINITAB output for sample statistics that have been discussed in this chapter is listed here.

$$\begin{array}{lll} b = 74.07 & a = 5.9772 & \text{SSR} = 86.404 \\ \text{SSE} = 77.414 & \text{MSE} = 86.404 & \text{MSR} = 4.301 \\ S_e = 2.074 & R^2 = .527 & \bar{R}^2 = .501. \end{array}$$

Other sample statistic outputs in Fig. 13.9 will be discussed in the next chapter.

Example 13.4 The Regression Relationship Between Number of Cars and Size of Household. Say we have random samples of 10 households showing the numbers of cars per household listed in Table 13.8. From Table 13.8, we can obtain the following statistics:

$$\bar{y} = 2.4 \qquad \bar{x} = 3.6 \qquad \Sigma x_i y_i = 99 \qquad \Sigma x_i^2 = 150$$
$$\Sigma y_i^2 = 68 \qquad \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 12.6$$
$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 20.4 \qquad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 10.4$$

Table 13.8 Numbers of cars per household

Household	Cars, y	People, x
1	4	6
2	1	2
3	3	4
4	2	3
5	2	4
6	3	4
7	4	6
8	1	3
9	2	2
10	2	2
Total	24	36

```
MTB > READ C1 C2
DATA> 4 6
DATA> 1 2
DATA> 3 4
DATA> 2 3
DATA> 2 4
DATA> 3 4
DATA> 4 6
DATA> 1 3
DATA> 2 2
DATA> 2 2
DATA> END
     10 rows read.
MTB > REGRESS C1 1 C2;
SUBC> DW.

Regression Analysis


The regression equation is
C1 = 0.176 + 0.618 C2

Predictor       Coef        StDev           T        P
Constant      0.1765       0.4905        0.36    0.728
C2            0.6176       0.1266        4.88    0.000

S = 0.5720      R-Sq = 74.8%     R-Sq(adj) = 71.7%

Analysis of Variance

Source        DF           SS           MS          F        P
Regression     1       7.7824       7.7824      23.78    0.000
Error          8       2.6176       0.3272
Total          9      10.4000

Durbin-Watson statistic = 2.44
```

Fig. 13.10 MINITAB output for Example 13.4

Following Eq. 13.15, we can estimate the regression slope as

$$b = \frac{\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{10} (x_i - \bar{x})^2} = \frac{12.6}{20.4} = .6176 = .62$$

This implies that, on the average, the number of cars for each household increases by approximately .62 when the number of people in the household increases by 1. Equation 13.14 yields an estimate of the intercept.

$$a = \frac{24}{10} - .62\left(\frac{36}{10}\right) = .168$$

The estimated regression line is

$$\hat{y} = .168 + .62x$$

From Eqs. 13.34 and 13.35, we can estimate R^2 and s_e as

$$R^2 = (.6176)^2\left(\frac{20.4}{10.4}\right) = .7482$$

$$s_e = \sqrt{[10.4 - (.6176)^2(20.4)]/8}$$

$$= .5721$$

Other related statistical analysis will be done in Example 14.1.

The MINITAB output of Example 13.4 is presented in Fig. 13.10, which displays most of the results we have calculated in this example. Some of the estimates of this output will be investigated in the next chapter.

13.7 Summary

In this chapter, we discussed the basic concepts of simple linear regression and the correlation coefficient. Both population and sample regression lines were defined. The least-squares method of estimating the intercept and slope of a regression line were also discussed. Coefficient of determination of a regression analysis was defined. In the next chapter, the ideas and analyses introduced in this chapter will be used for further analysis. And applications of simple regression in business and economic decisions will be explored.

Questions and Problems

1. Discuss the standard assumptions for linear regression analysis.
2. A study by the New York/New Jersey Port Authority on the effects of train ticket prices on the number of passengers produced the following results:

Ticket price	Passengers/hour
$6.00	500
$6.50	490
$7.00	475
$7.50	450
$8.00	400
$8.50	350

(a) Which variable should be the independent variable and which the dependent variable?
(b) Plot the data.
(c) Use the method of least squares to estimate the slope and intercept.

3. A Department of Agriculture research team has investigated the relationship between the wheat harvest and the amount of fertilizer used.

Pounds of fertilizer per acre	20	30	40	50	60
Bushels of wheat per acre	100	111	120	135	145

(a) Plot the data.
(b) Use the method of least squares to estimate the slope and intercept.
(c) Predict the number of bushels of wheat that will be grown if 35 lb of fertilizer is used.

4. As vice president in charge of marketing, Bob Seller is interested in the relationship between dollars spent on advertising and the number of widgets his company sells. He has collected the following data on advertising dollars and numbers of widgets sold:

Advertising dollars (thousands)	10	15	25	70	100
Widgets sold (thousands)	100	120	145	250	400

(a) Which variable should be the dependent variable and which the independent variable?
(b) Plot the data.
(c) Use the method of least squares to estimate the slope and intercept.
(d) Predict the sale of widgets if $ 175,000 is spent on advertising.

5. Financial economists are often interested in measuring the relationship between the return on an individual stock and the return on the S&P 500. This model is usually referred to as the market model. Use the MINITAB program and the following rates of return for Ford stock and the S&P 500 in the table to:

(a) Plot the data. (Hint: Follow the procedure presented in Fig. 13.7.)
(b) Use the method of least squares to estimate the slope and intercept. (Hint: Follow the procedures presented in Fig. 13.8.)
(c) Calculate the standard error of the estimates.
(d) Calculate the coefficient of determination.

Year	Ford	S&P 500
70	.4260	.0010
71	.2933	.1080
72	.1717	.1557

(continued)

(continued)

Year	Ford	S&P 500
73	–.4512	–.1737
74	–.0968	–.2964
75	.3960	.3149
76	.4614	.1918
77	–.2067	–.1153
78	–.0026	.0105
79	–.1479	.1228
80	–.2938	.2586
81	–.1025	–.0994
82	1.3212	.1549
83	.1286	.1706
84	.0980	.0115
85	.3237	.2633
86	.0081	.1462
87	.3961	.0203
88	–.2995	.1240
89	–.0767	.2725
90	–.3209	–.0656

6. Explain whether you would expect a positive relationship, a negative relationship, or no relationship to exist for the following pairs of data. If you think there is a relationship, identify the dependent variable.

   (a) The height of a mother and that of her son
   (b) The income and age of female accountants
   (c) The height and weight of a gorilla
   (d) The cost of a car and the cost of insuring that car
   (e) The time it takes a woman to run a marathon and the number of hours she spends training

7. Mary Jones, a professor of statistics, has collected the following sample of hours spent studying for her course and grades received on the midterm exam.

Sampled student	1	2	3	4	5	6	7	8	9
Hours of study	22	18	30	22	29	35	18	21	40
Exam grade	63	59	85	70	90	93	72	75	98

   (a) Plot the data.
   (b) Use the method of least squares to estimate α and β.
   (c) Use the regression equation to predict the grade of a student who spent 25 hours studying.

8. An English professor has estimated the following relationship between English SAT scores (x) and score in the freshman English course (y).

$$\hat{y} = 30 + .12x \quad R^2 = .35$$
$$(.79)\ (.05)$$

where standard deviations are shown in parentheses. The average SAT score for these students was 550.

(a) What is the students' average score in this course?
(b) Use the regression equation to predict the English course score for a student with an English SAT score of 400, 500, 600, 700, and 800.
(c) If there is a 50-point difference in the SAT scores of two students at this school, what is the predicted difference in their course scores?

9. Elmore Truesdale, vice president in charge of strategic pricing, is trying to find the relationship between the price of widgets and the quantity of widgets sold. Mr. Truesdale has collected the following data:

Price	\$12.50	12.00	11.50
Widgets sold (thousands)	125	135	140
Price	\$11.00	10.50	10.00
Widgets sold (thousands)	148	170	185

(a) Plot the data.
(b) Use the method of least squares to estimate α and β.
(c) Use the regression model to predict how many widgets would be sold if the price of widgets were \$9.80.

10. The following table gives data on personal consumption C and disposable income Y^d in the United States. Use the MINITAB or SAS program to answer the following.

Year	C	Y^d
1976	1803.9	2001.0
1977	1883.8	2066.6
1978	1961.0	2167.4
1979	2004.4	2212.6
1980	2000.4	2214.3
1981	2024.2	2248.6
1982	2050.7	2261.5
1983	2145.9	2334.6
1984	2239.9	2468.4

(a) Plot the data, using C as the dependent variable.
(b) Use the method of least squares to calculate α and β.
(c) Interpret α and β.
(d) Calculate the standard error of the estimates.

11. The following table shows the annual rates of return for several assets and the rate of inflation.

Year	Common stocks	Corporate bonds	Treasury bonds	Rate of inflation
1967	24.0 %	−5.0 %	−9.2 %	3.0 %
1968	11.1	2.6	−.3	4.7
1969	−8.5	−8.1	−5.1	6.1
1970	4.0	18.4	12.1	5.5
1971	14.3	11.0	13.2	3.4
1972	19.0	7.3	5.7	3.4
1973	−14.7	1.1	−1.1	8.8
1974	−26.5	−3.1	4.4	12.2
1975	37.2	14.6	9.2	7.0
1976	23.8	18.6	16.8	4.8
1977	−7.2	1.7	−.7	6.8
1978	6.6	−.1	−1.2	9.0
1979	18.4	−4.2	−1.2	13.3
1980	32.4	−2.6	−4.0	12.4
1981	−4.9	−1.0	1.8	8.9
1982	21.4	43.8	40.3	3.9
1983	22.5	4.7	.7	3.8
1984	6.3	16.4	15.4	4.0
1985	32.2	30.9	31.0	3.8
1986	18.6	18.5	23.4	1.1

(a) Use the method of least squares to estimate the relationship between the rate of return on common stocks and the rate of inflation by using the MINITAB program.
(b) Do common stocks serve as a hedge against inflation?
(c) Repeat parts (a) and (b), using corporate bond returns.
(d) Repeat parts (a) and (b), using treasury bond returns.

12. Briefly explain the difference between a dependent variable and an independent variable in regression analysis.
13. What is causality? What is correlation? What is the relationship among casualty, correlation, and regression analysis?
14. Explain the difference between a population and a subpopulation.
15. What is a scatter diagram? Briefly explain the concept of a regression line in the context of a scatter diagram.
16. The *market model* equation in finance is

$$R_{j,t} = \alpha_j + \beta_j R_{m,t} + e_{j,t}$$

where
R_{ji} = return on stock j in month t
$R_{m,t}$ = return on the S&P 500 in month t
$e_{j,j}$ = error term

(a) What is the independent variable?
(b) What is the dependent variable?
(c) What are the regression coefficients?

17. Suppose you collect data on household consumption and income in the United States and estimate the regression equation C–500 + .8 Y, where C = consumption and K = income.

(a) Plot the relationship between consumption and income that this equation reflects.
(b) Explain the relationship between consumption and income.

18. Suppose you estimate the regression equation $y = 5 + .6x$

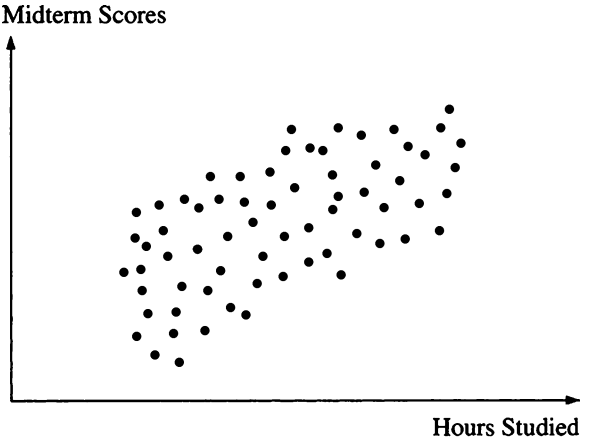
(a) What is the dependent variable?
(b) What is the independent variable?
(c) What is the intercept?
(d) What is the slope?

19. What is a sample? What is a population? Briefly explain how a sample can be used to estimate population parameters.
20. Briefly explain what we mean by the method of least squares.
21. You are given the following information on the heights and weights of 5 people:

Weight (pounds)	Height (inches)
180	72
165	66
130	62
220	78
110	60

(a) If you are interested in finding the relationship between height and weight, which variable should be the dependent variable?
(b) Use the method of least squares to estimate the slope and intercept.

22. Use the data given in question 21 and the results you got there to plot the regression line. Calculate the estimated values of y and the error from the regression.
23. The consumption function can be estimated by regressing private consumption on GNP. Use the data given in Table 2.2 and the MINITAB or SAS program to estimate the consumption function.
24. What do we mean when we say that estimates of α and β determined by the least-squares method are BLUE?
25. Briefly explain what we mean by a direct and by an inverse relationship.
26. The accompanying scatter diagram shows the numbers of hours several students studied and their midterm scores.



(a) Which is the dependent variable and which the independent variable?
(b) Is there a direct or an inverse relationship between hours studied and midterm score?
(c) Explain how we use regression analysis to estimate the relationship between hours studied and midterm score.

27. Suppose you are a safety consultant for the Department of Transportation. You are interested in the relationship between the number of miles a trucker drives per year and the number of accidents he or she has per year. You collect the following information from 6 truckers.

Trucker	Miles driven	Accidents
1	90,000	3
2	119,000	4
3	87,000	2
4	135,000	6
5	150,000	5
6	92,000	3

(a) Draw a scatter diagram showing the relationship between miles driven and number of accidents.
(b) Is there a direct or an inverse relationship between miles driven and number of accidents?

28. Use the method of least squares to estimate the intercept and slope of the regression line for the data given in question 27.
29. Use the data from question 27 and your results from question 28 to estimate the number of accidents for each trucker. Also calculate the errors from the regression line.
30. Suppose a labor economist at the Department of Labor estimates the following relationship between years of experience and earnings of accountants.

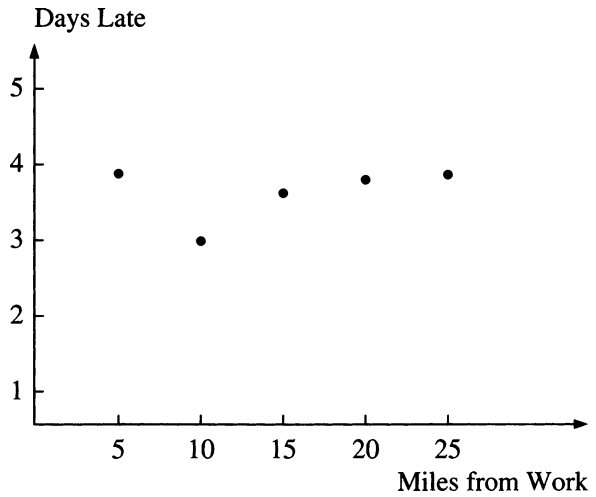
Earnings = 22,000 + 3,200(years of experience)
Estimate the earnings for the following 5 accountants:

Accountant	Years of experience
Bob	10.2
Mary	6.5
Sue	3.4
Ted	5.3
Anne	12.7

31. Now suppose the actual earnings of the 5 accountants in question 30 are as follows. Calculate the error from the regression.

Accountant	Actual earnings
Bob	$63,000
Mary	37,000
Sue	32,000
Ted	41,000
Anne	71,000

32. In order to determine whether a company should encourage its employees to live close to work, an efficiency expert collects data on the number of latenesses per month and the number of miles an employee lives from work. The relationship is given in the accompanying scatter diagram. From this scatter diagram, describe the relationship between miles from work and number of latenesses. Will the use of a regression line be helpful in estimating this relationship?

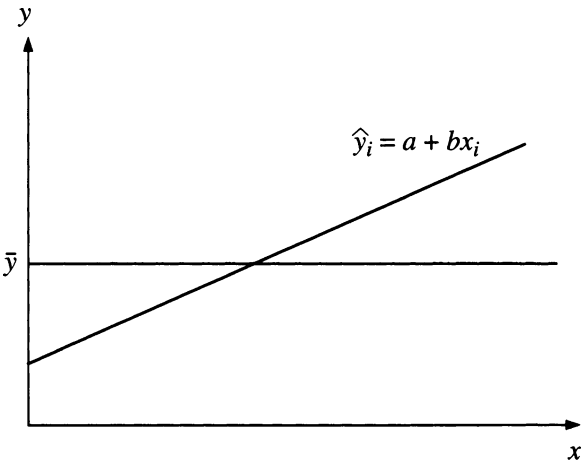


33. The manager of the Tow Time Auto Club would like to know the relationship between the age of a car and the number of service calls per year. He collects the following information:

x Car's age (years)	y Service calls per year
.5	0
1	2
2.5	1
3.5	5
4.2	8
5.6	7

(a) Draw a scatter diagram for these data.
(b) Use the method of least squares to estimate the parameters α and β.

34. Use the results from question 33 to predict the number of service calls for a car that is 3 years old and for one that is 6 years old.
35. Explain what we mean by the goodness of fit for a regression. Give two measures we can use to assess the goodness of fit.
36. Suppose you have the following information about number of dollars spent on advertising, X, and amount of car sales, Y: $\text{Cov}(X,Y) = 500$, $\text{Var}(X) = 250$, $\text{Var}(Y) = 1{,}000$, mean of $X = 1{,}000$, and mean of $Y = 2{,}500$. Use this information to estimate the parameters α and β.
37. Use the information given in question 36 to find the correlation between advertising dollars and sales.
38. Briefly define total deviation, unexplained deviation, and explained deviation.
39. Use the data and results from question 21 to calculate SSE, SSR, SST, and the coefficient of determination. Interpret what the coefficient tells us.
40. Use the data and results from question 21 to calculate the standard error of the residuals.
41. Use the data and results of questions 30 and 31 to calculate SSE, SSR, SST, and the coefficient of determination.
42. Calculate the standard error of the residual using the results from question 41.
43. Look at the scatter diagram given in question 32. Would you expect a regression on that data to produce a high or a low coefficient of determination?
44. Explain the difference between the coefficient of determination and the adjusted coefficient of determination. Which do you believe provides a better measure of the goodness of fit of a regression?
45. Briefly explain what the slope of a regression line tells us.
46. Look at the following graph and identify the explained error, the unexplained error, and the total error.



Use the MINITAB or SAS program to answer questions 47–51.

47. Use the data given in Table 2B.2 in Appendix B of Chap. 2 to estimate the relationship between GM stock's rate of return and the rate of return for the S&P 500. Calculate α, β, and the coefficient of determination.
48. Redo question 47, but this time, estimate the relationship between the rates of return for Ford and the S&P 500.
49. Use the data given in Table 2B.1 to estimate the regression for DPS regressed on EPS for General Motors. Calculate the coefficient of determination.
50. Use the data given in Table 2B.1 to estimate the regression for PPS regressed on EPS for Ford. Calculate the coefficient of determination.
51. Use the data given in Table 2B.1 to calculate the correlation coefficient between Ford's and GM's EPS.
52. Suppose you are interested in finding the relationship between bond prices and interest rates. You run a regression of bond prices against the prime lending rate and find that the slope of the regression line is negative. What does this tell you about the relationship between bond prices and interest rates? Is this relationship consistent with standard financial theory?
53. What type of correlation (positive, negative, or zero) would you expect from the following pairs of variables?

    (a) A company's earnings per share and its dividends per share
    (b) A company's earnings per share and its price per share
    (c) GM's EPS and the auto industry's average EPS
    (d) Education and salary of an employee
    (e) Advertising dollars spent and volume of sales
    (f) Bond prices and interest rates
    (g) The price charged for bread and the quantity of bread sold
    (h) The hem length of dresses in France and the value of the Dow Jones Industrial Average

54. What is the relationship between the correlation coefficient r and the coefficient of determination R^2?
55. Suppose you collect data for IBM's sales and dollars spent on advertising and then compute the following statistics:
    Cov(sales, advertising \$) = 22
    Var(sales) = 10
    Var(advertising \$) = 64
    Mean sales = 100
    Mean advertising \$ = 20

    (a) Compute the correlation coefficient between advertising dollars and sales.
    (b) Calculate the coefficient of determination that would result from a regression of sales on advertising dollars.
    (c) Calculate the regression parameters α and β.
56. Suppose you estimate a regression and compute SSE = 17.57 and SSR = 102.76. Calculate SST, R^2, and r by using this information.
57. The Department of Accounting at a university is interested in the relationship between SAT score and graduating grade point average (GPA). The accompanying table presents a summary of the data it has collected.

SAT, x	GPA y	xy
600	3.2	1920
420	2.5	1050
750	3.9	2925
650	3.6	2340
550	3.4	1870
680	3.7	2516
$\Sigma x = 3650$	$\Sigma y = 20.3$	$\Sigma xy = 12621$

x^2	y^2
360000	10.24
176400	6.25
562500	15.21
422500	12.96
302500	11.56
462400	13.69
$\Sigma x^2 = 2286300$	$\sum y^2 = 69.91$

    (a) Draw a scatter diagram for these data.
    (b) Estimate the regression parameters α and β.
58. Use the data given in question 57 to calculate r and R^2. Also use your regression results to estimate the graduating GPA for someone who scores 620 on the SAT.
59. The Department of Education at a university is interested in the relationship between the number of years of education and a person's salary. It collects the following information:

Person	Education (years)	Salary
1	8	$21,000
2	12	24,000
3	13	19,500
4	14	40,000
5	16	72,000

(a) Draw a scatter diagram for these data.
(b) Calculate the regression parameters α and β.

60. Use the data and your results from question 59 to estimate the earnings of someone with 15 years of education. Also compute r and R^2.
61. A market researcher is interested in who buys Fun Time Cereal. In order to analyze this problem, she collects data on the age of the consumer and how many boxes that person consumes each month.

Age	Boxes per month
8	6
10	8
16	5
22	4
35	2
45	0

(a) Compute the correlation between age and number of boxes of cereal consumed (and presumably purchased).
(b) Use the method of least squares to estimate α and β.

62. Use the information given in question 61 to calculate the standard error of the residual. Briefly explain how we can use the standard error of the residual as a measure of the goodness of fit.

Use the MINITAB or SAS program to answer questions 63–69.

63. Use the data given in question 23 of Chap. 2 to compute the correlation between the dollar/pound exchange rate and the dollar/yen exchange rate.
64. Use the data given in question 24 of Chap. 2 to compute the correlation coefficient between J&J's current ratio and the industry's.
65. Use the data given in question 24 of Chap. 2 to estimate the regression coefficients for a regression of J&J's current ratio on the pharmaceutical industry's current ratio.
66. Use your results from question 65 to compute the standard error of the estimate.
67. Repeat questions 64–66, using J&J's inventory turnover.
68. Repeat questions 64–66, using J&J's ROA.
69. Repeat questions 64–66, using J&J's price–earnings ratio.

70. You are given the following information on unemployment in the United States and in New Jersey (data from New Jersey Economic Indicators, March 1990): *Number unemployed* (*in thousands*)

Year	United States	New Jersey
1970	4,093	138
1971	5,016	172
1972	4,882	182
1973	4,365	180
1974	5,156	204
1975	7,929	334
1976	7,406	346
1977	6,991	317
1978	6,202	248
1979	6,137	247
1980	7,637	260
1981	8,273	263
1982	10,678	326
1983	10,717	288
1984	8,539	236
1985	8,312	217
1986	8,237	197
1987	7,425	160
1988	6,701	151
1989	6,528	163

If you are interested in the relationship between unemployment in the United States and in New Jersey, which unemployment figure should be your independent variable? Use the MINITAB program to estimate a model showing the relationship between unemployment in the United States and in New Jersey.

71. Use the information and MINITAB results from question 70 to compute the standard error of the estimate and the coefficient of determination.
72. Use the data given in question 70 and the MINITAB program to find the correlation coefficient between unemployment in the United States and in New Jersey for 1980–1989.
73. Consider the following table. Fill in the values missing from the table, using the least-squares method.

$$x \; y \; xy \; x^2 \; y^2 \; \hat{y} \; e \; e^2 \; (y - \bar{y})^2$$

x	y
5	50
7	35
9	25
11	20
13	15
15	10

74. Use your results in question 73 to find the coefficient of determination and the standard error of the estimate.
75. Suppose you estimate a regression and compute SST = 217.47 and SSR = 121.73. Use this information to calculate SSE, R^2, and r.
76. Suppose you estimate a regression and compute SST = 1017.17and SSE = 302.33. Use this information to calculate SSR, R^2, and r.
77. Suppose you are interested in finding the relationship between the monthly highest prices of a stock and risk-free interest rates. A regression of the highest stock prices against the risk-free interest rates shows that the slope of the regression line is positive. What does this tell you about the relationship between the monthly highest prices and risk-free interest rates?
78. Compute the correlation coefficient and the R^2 between the monthly highest prices of a stock (y, unit: dollar) and risk-free interest rates (x, unit: %).

x	y
1.70	12.5
1.74	12.1
1.75	11.9
1.91	17.4
2.17	24.8
2.06	19.3
2.04	19.5
1.93	18.2
1.88	17.6
1.92	19.6
1.90	25.2
1.91	31.3

79. Use the information given in Problem *78* to estimate the regression of the monthly highest stock price on the risk-free interest rate.
80. Use the estimated regression equation from Problem *79*. Interpret the results of this regression.
81. Use the estimated regression equation from Problem *79*. How much is the highest stock price expected to increase or decrease if the risk-free interest rate rises by 0.3?

Appendix 1: Derivation of Normal Equations and Optimal Portfolio Weights

In this appendix, we derive the normal equations that are used to obtain the least-squares estimates of population regression parameters. For convenience, we denote the function to be minimized as

$$F = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{13.36}$$

Because this function is to be minimized with respect to a and b, it is necessary to take the partial derivatives of F with respect to these two variables. The partial derivatives are

$$\frac{\partial F}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-1)$$
$$\frac{\partial F}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i)$$

Setting these partial derivatives equal to zero yields the following two normal equations:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$
$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \tag{13.37}$$

These are Eqs. 13.12 and 13.13 in the text. Note that setting the first partial equal to zero is identical to requiring that the sum of the residuals be zero because the term in parentheses is the residual $e_i = (y_i - a - bx_i)$.

Now we use the technique of deriving Eq. 13.37 to derive the optimal weight of a portfolio. Following Equation (6.29) in Chap. 6, the variance of rates of return for a portfolio is defined as

$$\text{Var}(Rp) = W_1^2\sigma_1^2 + W_2^2\sigma_2^2 + 2W_1W_2\,\sigma_{12}$$
$$(W_1 + W_2 = 1) \tag{13.38}$$

where W_1 and W_2 represent percentage money invested in security 1 and security 2, respectively; σ_1^2 = variance of rates of return for security 1, σ_2^2 = variance of rates of return for security 2, and σ_{12} = covariance between the rates of return for security 1 and the rates of return for security 2.

If the objective of the investor is to minimize the variance of a portfolio, then the optimal weights of a two-security portfolio can be obtained by taking partial derivatives of $Var(Rp)$ with respect to the variance of W_1 and $W_2 = 1 - W_1$ as:

$$\frac{\partial \text{Var}(Rp)}{\partial W_1} = 2W_1\sigma_1^2 - 2(1 - W_1)\sigma_2^2 + 2(1 - W_1)\sigma_{12}$$

Setting this partial derivative to zero and solving for W_1, we obtain

$$W_1 = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} \tag{13.39}$$

$$W_2 = 1 - W_1 = \frac{\sigma_1^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} \tag{13.40}$$

Substituting the data of Example 6.21 in Chap. 6 into these two equations, we can estimate W_1 and W_2 as[11]

$$W_1 = \frac{.008 + .00375}{.00625 + .008 + 2(.00375)} = .5402$$
$$W_2 = 1 - .542 = .4598$$

Appendix 2: The Derivation of Equation 13.20

The left-hand side of Eq. 13.20 can be written as

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{13.41}$$

In addition, we know that

$$\begin{aligned}\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} [a + bx_i - (a + b\bar{x})][y_i - \hat{y}_i] \\ &= b\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \hat{y}_i) \\ &= -b\bar{x}\sum_{i=1}^{n} (y_i - \hat{y}) + b\sum_{i=1}^{n} (y_i - \hat{y}_i)(x_i)\end{aligned}$$

Because assumptions C and A discussed in Sect. 13.4 imply that

$$\sum_{i=1}^{n} (y_i - \hat{y}) = \sum_{i=1}^{n} e_i = 0 \text{ and } \sum_{i=1}^{n} (y_i - \hat{y}_i)(x_i) = \sum_{i=1}^{n} e_i x_i = 0$$

[11] The weights obtained here do not consider the information of the expected rates of return for both stock A and stock B. The formula of estimating the optimal weights in terms of both variances and expected rates of return can be found in Chap. 8 of Cheng F. Lee *et al*. (1990), *Security Analysis and Portfolio Management* (Glenview, Ill.: Scott Foresman/Little, Brown).

Hence,

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0+0=0 \tag{13.42}$$

Substituting Eq. 13.42 into Eq. 13.41, we obtain Eq. 13.20.

Appendix 3: The Bivariate Normal Density Function

In correlation analysis, we assume a population where both X and Y vary jointly. It is called a joint distribution of two variables. If both X and Y are normally distributed, then we call this known distribution a *bivariate normal distribution*.

Following Appendix 1 of chap. 7, we can define the probability density function (PDF) of the normally distributed random variables X and Y as

$$f(X)=\frac{1}{\sigma_X\sqrt{2\pi}}\exp\left[\frac{-(\mathrm{X}-\mu_\mathrm{X})^2}{2\sigma_\mathrm{X}^2}\right],-\infty<\mathrm{X}<\infty \tag{13.43}$$

$$f(Y)=\frac{1}{\sigma_Y\sqrt{2\pi}}\exp\left[\frac{-(\mathrm{Y}-\mu_\mathrm{Y})}{2\sigma_\mathrm{Y}^2}\right],-\infty<\mathrm{Y}<\infty \tag{13.44}$$

where μ_X and μ_Y are population means for X and Y, respectively; σ_X and σ_Y are population standard deviations of X and Y, respectively; $\pi=3.1416$; and exp represents the exponential function.

If ρ represents the population correlation between X and Y, then the PDF of the bivariate normal distribution can be defined as

$$f(X,Y)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp(-q/2),\quad -\infty<X<\infty,-\infty<Y<\infty \tag{13.45}$$

where $\sigma_X>0$, $\sigma_Y>0$, and $-1<\rho<1$,

$$q=\frac{1}{1-\rho^2}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2-2\rho\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)+\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right]$$

It can be shown that the conditional mean of Y, given X, is linear in x and given by

$$E(Y|X) = \mu_Y + \rho\left\{\frac{\sigma_Y}{\sigma_X}\right\}(X - \mu_X) \tag{13.46}$$

It is also clear that given X, we can define the conditional variance of Y as

$$\sigma^2(Y|X) = \sigma_Y^2\left(1 - \rho^2\right) \tag{13.47}$$

Equation 13.46 can be regarded as describing the population linear regression line. For example, if we have a bivariate normal distribution of heights of brothers and sisters, we can see that they vary together and there is no cause-and-effect relationship. Accordingly, a linear regression in terms of the bivariate normal distribution variable is treated as though there were a two-way relationship instead of an existing causal relationship. It should be noted that regression implies a causal relationship only under a *prediction* case.

Equation 13.45 represents a joint PDF for X and Y. If $p = 0$, then Eq. 13.45 becomes

$$f(X, Y) = f(X)f(Y) \tag{13.48}$$

This implies that the joint PDF of X and Y is equal to the PDF of X times the PDF of Y. We also know that both X and Y are normally distributed. Therefore, X is independent of Y.

Example 13.5 Using a Mathematics Aptitude Test to Predict Grade in Statistics. Let X and Y represent scores in a mathematics aptitude test and numerical grade in elementary statistics, respectively. In addition, we assume that the parameters in Eq. 13.45 are

$$\mu_X = 550 \quad \sigma_X = 40 \quad \mu_Y = 80 \quad \sigma_Y = 4 \quad \rho = .7$$

Substituting this information into Equations 13.46 and 13.47, respectively, we obtain

$$\begin{aligned} E(Y|X) &= 80 + .7(4/40)(X - 550) \\ &= 41.5 + .07X \end{aligned} \tag{13.49}$$

$$\sigma^2(Y|X) = (16)(1 - .49) = 8.16 \tag{13.50}$$

If we know nothing about the aptitude test score of a particular student (say, John), we have to use the distribution of Y to predict his elementary statistics grade.

$$95\%\,\text{interval} = 80 \pm (1.96)(4) = 80 \pm 7.84$$

That is, we predict with 95 % probability that John's grade will fall between 87.84 and 72.16.

Alternatively, suppose we know that John's mathematics aptitude score is 650. In this case, we can use Eqs. 13.49 and 13.50 to predict John's grade in elementary statistics.

$$E(Y|X = 650) = 41.5 + (.07)(650) = 87$$

and

$$\sigma^2(Y|X = 650) = 8.16$$

We can now base our interval on a normal probability distribution with a mean of 87 and a standard deviation of 2.86.

$$95\%\,\text{interval} = 87 \pm (1.96)(2.86) = 87 \pm 5.61$$

That is, we predict with 95 percent probability that John's grade will fall between 92.61 and 81.39.

Two things have happened to this interval. First, the center has shifted upward to take into account the fact that John's mathematics aptitude score is above average. Second, the width of the interval has been narrowed from $87.84-72.16 = 15.68$ grade points to $92.61-81.39 = 11.22$ grade points. In this sense, the information about John's mathematics aptitude score has made us less uncertain about his grade in statistics. This issue is discussed in further detail in Sect. 14.4 in the next chapter.

Appendix 4: American Call Option and the Bivariate Normal CDF

The call option pricing model discussed in Appendix 2 of Chap. 6 and Appendices 2 and 3 of Chap. 7 is derived in terms of an option contract which can be exercised only on the expiration date. This kind of option is called *European call*. If the contract of a call option can be exercised at any time of the option's contract period, then this kind of call option is called *American call*.

When a stock pays a dividend, the *American call* is more complex. *The American call* is with one known dividend payment. The valuation equation can be defined as[12]

$$\begin{aligned} C(S,T,X) = {} & S^x\left[N_1(b_1) + N_2\left(a_1, -b_1; -\sqrt{t/T}\right)\right] \\ & - Xe^{-rT}\left[N_1(b_2)e^{r(T-t)} + N_2\left(a_2, -b_2; -\sqrt{t/T}\right)\right] \\ & + De^{-rT}N_1(b_2) \end{aligned} \tag{13.51}$$

where

$$a_1 = \frac{\ln\left(\frac{S^x}{X}\right) + \left(r + \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}, \quad a_2 = a_1 - \sigma\sqrt{T} \tag{13.52}$$

$$b_1 = \frac{\ln\left(S^x/S_t^*\right) + \left(r + \frac{1}{2}\sigma^2\right)t}{\sigma\sqrt{t}}, \quad b_2 = b_1 - \sigma\sqrt{t} \tag{13.53}$$

$$S^x = S - De^{-rT} \tag{13.54}$$

S^X represents the correct stock net price of the present value of the promised dividend per share (D). t represents the time dividend to be paid.

S_t^* is the ex-dividend stock price for the American call option which

$$C\left(S_t^*, T-t, X\right) = S_t^* + D - X \tag{13.55}$$

S, X, r, σ^2, T have been defined in Appendix 3 of Chap. 7.

Both $N_1(b_1)$ and $N_1(b_2)$ are cumulative univariate normal density function. N_2 (a, b; p) is the cumulative bivariate normal density function with upper integral limits, a and b, and correlation coefficient, $\rho = -\sqrt{t/T}$.

American call option on a non-dividend-paying stock will never optimally be exercised prior to expiration. Therefore, if there exist no dividend payments, Eqs. 13.51, 13.52, 13.53 will reduce to the valuation Equation of the European Option with no dividend payment as defined in Eq. 7.35 of Appendix 2 of Chap. 7.

In Appendices 1 and 2 of Chap. 7, we have shown how the cumulative univariate normal density function can be used to evaluate the European call option. In this appendix, we found that if a common stock pays a discrete dividend during the option's life, the American call option valuation equation requires the evaluation of a cumulative bivariate normal density function. While there are many available

[12] This equation is based upon Whaley, Robert E. (1981), "On the Valuation of American Call Options on Stocks With Known Dividends," *Journal of Financial Economics* 9, 207–211.

approximations for the cumulative bivariate normal distribution, the approximation provided here relies on Gaussian quadratures. The approach is straightforward and efficient, and its maximum absolute error is .00000055.

Following Eq. 13.45 in Appendix 3, the probability that x' is less than a and that y' is less than b for the standardized cumulative bivariate normal distribution

$$P(X' < a, Y' < b) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{a} \int_{-\infty}^{b} \exp\left[\frac{2x'^2 - 2\rho x'y' + y'^2}{2(1-\rho^2)}\right] dx'dy'$$

where $x' = \frac{x-\mu_X}{\sigma_X}, y' \frac{y-\mu_Y}{\sigma_Y}$ and ρ is the correlation between the random variables x' and y'.

The first step in the approximation of the bivariate normal probability $N_2(a,b;\rho)$ is as follows:

$$\phi(a,b;\rho) \approx .31830989\sqrt{1-\rho^2} \sum_{i=1}^{5} \sum_{j=1}^{5} w_i w_j f\left(x_i', x_j'\right), \tag{13.56}$$

where

$$f\left(x_i', x_j'\right) = \exp\left[a_1\left(2x_i' - a_1\right) + b_1\left(2x_j' - b_1\right) + 2\rho\left(x_i' - a_1\right)\left(x_j' - b_1\right)\right],$$

The pairs of weights (w) and corresponding abscissa values (x') are[13]

i, j	w	x'
1	.24840615	.10024215
2	.39233107	.48281397
3	.21141819	1.0609498
4	.033246660	1.7797294
5	.00082485334	2.6697604

and the coefficients a_1 and b_1 are computed using

$$a_1 = \frac{a}{\sqrt{2(1-\rho^2)}} \quad \text{and} \quad b_1 = \frac{b}{\sqrt{2(1-\rho^2)}}$$

The second step in the approximation involves computing the product $ab\rho$.

[13] This portion is based upon Appendix 13.1 of Hans R. Stoll and Robert E. Whaley (1993), *Futures and Options* (South Western Publishing, Cincinnati).

If $ab\rho \leq 0$, compute the bivariate normal probability, N_2 (a,b;ρ), using the following rules:

1. If $a \leq 0, b \leq 0,$ and $\rho \leq 0$, then $N_2(a,b;\rho) = \phi(a,b;\rho)$.
2. If $a \leq 0, b \geq 0,$ and $\rho > 0$, then $N_2(a,b;\rho) = N_1(a) - \phi(a,-b;-\rho)$.
3. If $a \geq 0, b \leq 0,$ and $\rho > 0$, then $N_2(a,b;\rho) = N_1(b) - \phi(-a,b;-\rho)$.
4. If $a \geq 0, b \geq 0,$ and $\rho \leq 0$, then $N_2(a,b;\rho) = N_1(a) + N_1(b) - 1 + \phi(-a,-b;\rho)$.

(13.57)

If $ab\rho > 0$, compute the bivariate normal probability, N_2 (a,b; ρ), as

$$N_2(a,b;\rho) = N_2(a,0;\rho_{ab}) + N_2(b,0;\rho_{ba}) - \delta \tag{13.58}$$

where the values of N_2 (•) on the right-hand side are computed from the rules for $ab\rho \leq 0$,

$$\rho_{ab} = \frac{(\rho a - b)\text{Sgn}(a)}{\sqrt{a^2 - 2\rho ab + b^2}}, \qquad \rho_{ba} = \frac{(\rho b - a)\text{Sgn}(b)}{\sqrt{a^2 - 2\rho ab + b^2}}$$
$$\delta = \frac{1 - \text{Sgn}(a) \times \text{Sgn}(b)}{4}$$

and

$$\text{S}gn(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

$N_1(d)$ is the cumulative univariate normal probability.

Example 13.6 Valuating American Option. An *American call* option whose exercise price is \$45 has an expiration time of 90 days. Assume the risk-free rate of interest is 8 percent annually, the underlying price is \$50, the standard deviation of the rate of return of the stock is 20 percent, and the stock pays a dividend of \$1.5 in exactly 50 days, (a) what is the *European call* value? (b) Can the early exercise be predicted? (c) What is the value of the *American call?*

(a) The current stock net price of the present value of the promised dividend is

$$s^x = 50 - (1.5)e^{-0.8\left(\frac{50}{365}\right)} = 48.516$$

Following Equation 7B.2, the European call value can be calculated as

$$C = (48.516)N(d_1) - (45)\left(e^{-0.8\left(\frac{90}{365}\right)}\right)N(d_2)$$

where

$$\begin{aligned} d_1 &= \frac{[\text{In}(48.516/45) + (.08 + .5(.20)^2)(90/365)]}{.20\sqrt{90/365}} \\ &= \frac{.075 + .025}{.099} \\ &= 1.010 \\ d_2 &= 1.010 - .099 = .911 \end{aligned}$$

From Table A.l, we obtain

$$\begin{aligned} N(1.010) &= .5 + .3438 = .8438 \\ N(.911) &= .5 + .3186 = .8186 \end{aligned}$$

So the European call value is

$$\begin{aligned} C &= (48.516)(.8438) - 45(.980)(.8186) \\ &= 4.8375 \end{aligned}$$

(b) The present value of the interest income that would be earned by deferring exercise until expiration is

$$\begin{aligned} X(1 - e^{-r(T-t)}) &= (45)[1 - e^{-.08(90\text{-}50)/365}] \\ &= 45[1 - .991] \\ &= .405 \end{aligned}$$

Since $d = 1.5 > .405$, therefore, the early exercise is not precluded.

(c) The value of the American call is now calculated as

$$\begin{aligned} C = (48.208) &\left[N_1(b_1) + N_2\left(a_1, -b_1; -\sqrt{50/90}\right)\right] \\ &- (45)e^{-.08(90/365)}\left[N_1(b_2)e^{.08(40/365)} + N_2\left(a_2, -b_2; -\sqrt{50/90}\right]\right. \\ &+ 2e^{-.08(50/365)}N_1(b_2) \end{aligned} \tag{13.59}$$

since both b_1 and b_2 depend on the critical ex-dividend stock price S_t^*, which can be determined by

$$C\left(S_t^*, 40/365; 45\right) = S_t^* + 1.5 - 45$$

By using trial and error, we find that $S_t^* = 44.756$. An Excel program used to calculate this value is presented in Fig. 13.11.

Caculation of S_t^*(critical ex-dividend stock price)	
S_t^*(critical ex-dividend stock price)=	44.7557142157122
X(exercise price of option)=	45
r(risk-free interest rate)=	0.08
σ(volatility of stock)=	0.2
T-t(expiration date - excise date)=	=(90-50)/365
d1=	=(LN(C3/C4)+(C5+C6^2/2)*(C7))/(C6*SQRT(C7))
d2=	=(LN(C3/C4)+(C5-C6^2/2)*(C7))/(C6*SQRT(C7))
D(divident)=	1.5
c(value of European call option to buy one share)=	=C3*NORMSDIST(C8)-C4*EXP(-C5*C7)*NORMSDIST(C9)
p(value of European put option to sell one share)=	=C4*EXP(-C5*C7)*NORMSDIST(-C9)-C3*NORMSDIST(-C8)
$c(S_t^*,T\text{-}t;X)\text{-}S_t^*\text{-}D+X$=	=C12-C3-C10+C4

Caculation of S_t^*(critical ex-dividend stock price)		
S_t^*(critical ex-dividend stock price)=	$ 49.824	$ 44.756
X(exercise price of option)=	$ 50.000	$ 45.000
r(risk-free interest rate)=	0.1	0.08
σ(volatility of stock)=	0.3	0.2
T-t(expiration date - excise date)=	0.0822	0.1096
d1=	0.0977	0.0833
d2=	0.0117	0.0171
D(divident)=	$ 2.000	$ 1.500
c(value of European call option to buy one share)=	$ 1.82	$ 1.26
p(value of European put option to sell one share)=	$ 1.59	$ 1.11
$c(S_t^*,T\text{-}t;X)\text{-}S_t^*\text{-}D+X$=	0	0

Caculation of S_t^*(critical ex-dividend stock price)	
S_t^*(critical ex-dividend stock price)=	49.8244471377206
X(exercise price of option)=	50
r(risk-free interest rate)=	0.1
σ(volatility of stock)=	0.3
T-t(expiration date - excise date)=	=(90-60)/365
d1=	=(LN(B3/B4)+(B5+B6^2/2)*(B7))/(B6*SQRT(B7))
d2=	=(LN(B3/B4)+(B5-B6^2/2)*(B7))/(B6*SQRT(B7))
D(divident)=	2
c(value of European call option to buy one share)=	=B3*NORMSDIST(B8)-B4*EXP(-B5*B7)*NORMSDIST(B9)
p(value of European put option to sell one share)=	=B4*EXP(-B5*B7)*NORMSDIST(-B9)-B3*NORMSDIST(-B8)
$c(S_t^*,T\text{-}t;X)\text{-}S_t^*\text{-}D+X$=	=B12-B3-B10+B4

Fig. 13.11 Microsoft Excel program for calculating S_t^*

Substituting $S^x = 48.208$, $X = \$45$, and S_t^* into Eqs. 13.52 and 13.53, we can calculate a_1, a_2, b_1, and b_2 as follows:

$$
\begin{aligned}
a_1 &= d_1 = 1.010 \\
a_2 &= d_2 = .911 \\
b_1 &= \frac{\text{In}\left(\frac{48.516}{44.756}\right) + \left[.08 + \frac{1}{2}(.20)^2\right]\left(\frac{50}{365}\right)}{(.20)\sqrt{50/365}} \\
&= \frac{.0807 + .0137}{.0740}\, 1.2757 \\
b_2 &= 1.2757 - .0740 = 1.2017
\end{aligned}
$$

In addition, we also know $\rho = \sqrt{\frac{50}{90}} = .7454$.

From the above information, we now calculate the related normal probability as follows:

Using Equation 7A.9 in Appendix 7A, we obtain

$$
\begin{aligned}
N_1(b_1) &= N_1(1.2757) = .8988 \\
N_1(b_2) &= N(1.2017) = .8851
\end{aligned}
$$

Following Eq. 13.58, we now calculate the values of $N_2(1.010, -1.2757; -.7454)$ and 7 $N_2(.911, -1.2017; -.7454)$ as follows:

Since $ab\rho > 0$ for both cumulative bivariate normal density function, therefore, we can use Eq. 13.58 to calculate the value of both $N_2(a,b,\rho)$ as follows:

$$
\rho_{ab} = \frac{[(-.7454)(1.010) + 1.2757](1)}{\sqrt{(1.010)^2 - 2(-.7454)(1.010)(-1.2757) + (-1.2757)^2}} = .6133
$$

$$
\rho_{ba} = \frac{[(-.7454)(-1.2757) - 1.010](-1)}{\sqrt{(1.010)^2 - 2(-.7454)(1.010)(-1.2757) + (-1.2757)^2}} = .0693
$$

$$
\delta = \frac{1 - (1)(-1)}{4} = \frac{1}{2}.
$$

$$
\begin{aligned}
&N_2(1.010, -1.2757; -.7454) = N_2(1.010, 0, .6133) + N_2(-1.2757, 0; 0693) - .5 \\
&= N_1(0) + N_1(-1.2757) - \phi(-1.010, 0; -.6133) - \phi(-1.2757, 0; -.0693) - .5 \\
&= .5 + .1010 - .0202 - .0456 - .5 = .0352
\end{aligned}
$$

Using Microsoft Excel programs presented in Figs. 13.12 and 13.13, we obtain

```
Module1 - 1

Option Explicit   ' Force explicit variable declaration.

Public Function Bivarncdf(a As Double, b As Double, rho As Double) As Double

   Dim rho_ab As Double, rho_ba As Double
   Dim delta As Double

   If (a * b * rho) <= 0 Then

      If (a <= 0 And b <= 0 And rho <= 0) Then

         Bivarncdf = Phi(a, b, rho)

      End If

      If (a <= 0 And b >= 0 And rho > 0) Then

         Bivarncdf = Application.WorksheetFunction.NormSDist(a) - Phi(a, -b, -rho)

      End If

      If (a >= 0 And b <= 0 And rho > 0) Then

         Bivarncdf = Application.WorksheetFunction.NormSDist(b) - Phi(-a, b, -rho)

      End If

     If (a <= 0 And b >= 0 And rho > 0) Then

        Bivarncdf = Application.WorksheetFunction.NormSDist(a) - Phi(a, -b, -rho)

     End If

     If (a >= 0 And b <= 0 And rho > 0) Then

        Bivarncdf = Application.WorksheetFunction.NormSDist(b) - Phi(-a, b, -rho)

     End If

     If (a >= 0 And b >= 0 And rho <= 0) Then

        Bivarncdf = Application.WorksheetFunction.NormSDist(a) + Application.WorksheetFunction.
NormSDist(b) - 1 + Phi(-a, -b, rho)

     End If

   Else

      rho_ab = ((rho * a - b) * IIf(a >= 0, 1, -1)) / Sqr(a ^ 2 - 2 * rho * a * b + b ^ 2)
      rho_ba = ((rho * b - a) * IIf(b >= 0, 1, -1)) / Sqr(a ^ 2 - 2 * rho * a * b + b ^ 2)
      delta = (1 - IIf(a >= 0, 1, -1) * IIf(b >= 0, 1, -1)) / 4

      Bivarncdf = Bivarncdf(a, 0, rho_ab) + Bivarncdf(b, 0, rho_ba) - delta

   End If

End Function

Public Function Phi(a As Double, b As Double, rho As Double) As Double

   Dim a1 As Double, b1 As Double
   Dim w(5) As Double, x(5) As Double
   Dim i As Integer, j As Integer
   Dim doublesum As Double
```

Fig. 13.12 (continued)

```
    a1 = a / Sqr(2 * (1 - rho ^ 2))
    b1 = b / Sqr(2 * (1 - rho ^ 2))

    w(1) = 0.24840615
    w(2) = 0.39233107
    w(3) = 0.21141819
    w(4) = 0.03324666
    w(5) = 0.00082485334

    x(1) = 0.10024215
    x(2) = 0.48281397
    x(3) = 1.0609498
    x(4) = 1.7797294
    x(5) = 2.6697604

    doublesum = 0

Module1 - 2

    For i = 1 To 5

        For j = 1 To 5

        doublesum = doublesum + w(i) * w(j) * Exp(a1 * (2 * x(i) - a1) + b1 * (2 * x(j) - b1) + 2
* rho * (x(i) - a1) * (x(j) - b1))

        Next j

    Next i

    Phi = 0.31830989 * Sqr(1 - rho ^ 2) * doublesum

End Function
```

Fig. 13.12 Microsoft Excel program for calculating function Phi (ϕ)

$$\begin{aligned}
&\phi(-1.010,\ 0;\ .6133) = .0202\\
&\phi(-1257,\ 0; -.0693) = .0460\\
&N_2(1.010, -1.2757; -.7454) = 0.0350\\
&\phi(-.911,\ 0; -.6559) = .0218\\
&\phi(-1.2017,\ 0; -.0143) = .0563\\
&N_2(.911, -1.2017; -.7454) = 0.0368
\end{aligned}$$

Substituting the related information into Eq. 13.59, we obtain

$$\begin{aligned}
C &= (48.208)[.8988 + .0350]\\
&\quad - (45)e^{-.08(90/365)}\left[(.8851)e^{-.08(40/365)} + .0368\right]\\
&\quad + 2e^{-.08(50/365)}(.8851)\\
&= \$\,5.603
\end{aligned}$$

All related results are presented in column C of Fig. 13.13.[14]

[14] Results of column B are a different set of data. It is good exercise for students to try them.

	A	B
1	Option Pricing Calculation	
2		
3	S(current stock price)=	$ 50.000
4	S_t^*(critical ex-dividend stock price)=	$ 44.756
5	S^x(current stock price NPV of promised dividend)=	$ 48.516
6	X(exercise price of option)=	$ 45.000
7	r(risk-free interest rate)=	0.080
8	σ(volatility of stock)=	0.200
9	T(expiration date)=	$ 0.247
10	t(excise date)=	$ 0.137
11	D(Dividend)=	$ 1.500
12	d1(non-dividend-paying)=	1.3092
13	d2(non-dividend-paying)=	1.2099
14	d1*(critical ex-dividend stock price)=	0.0833
15	d2*(critical ex-dividend stock price)=	0.0171
16	d1(dividend-paying)=	1.0059
17	d2(dividend-paying)=	0.9066
18	a1=	1.0059
19	a2=	0.9066
20	b1=	1.2749
21	b2=	1.2009
22		
23	c(St*,T-t;X)=	$ 1.26
24	$c(S_t^*,T-t;X)-S_t^*-D+X$=	0
25		
26	$N_1(a_1)$=	0.8428
27	$N_1(a_2)$=	0.8177
28	$N_1(b_1)$=	0.8988
29	$N_1(b_2)$=	0.8851
30	$N_1(-b_1)$=	0.1012
31	$N_1(-b_2)$=	0.1149
32	ρ=	-0.7454
33	$a=a_1;b=-b_1$	
34	$\phi(a,-b;-\rho)$=	0.0460
35	$\phi(-a,b;-\rho)$=	0.0202
36	ρ_{ab}=	0.6166
37	ρ_{ba}=	0.0653
38	$N_2(a,0;\rho_{ab})$=	0.4798
39	$N_2(b,0;\rho_{ba})$=	0.0552
40	δ=	0.5000
41	$a=a_2;b=-b_2$	
42	$\phi(a,-b;-\rho)$=	0.0563
43	$\phi(-a,b;-\rho)$=	0.0218
44	ρ_{ab}=	0.6559
45	ρ_{ba}=	0.0143
46	$N_2(a,0;\rho_{ab})$=	0.4782
47	$N_2(b,0;\rho_{ba})$=	0.0586
48	δ=	0.5000
49		
50	$N_2(a_1,-b_1;\rho)$=	0.0350
51	$N_2(a_2,-b_2;\rho)$=	0.0368
52		
53	c(value of European call option to buy one share)=	$ 4.811
54	p(value of European put option to sell one share)=	$ 0.416
55	C(value of American call option to buy one share)=	$ 5.603

Fig. 13.13 (continued)

	A
1	Option Pricing Calculation
2	
3	S(current stock price)=
4	S_t^*(critical ex-dividend stock price)=
5	S^x(current stock price NPV of promised dividend)=
6	X(exercise price of option)=
7	r(risk-free interest rate)=
8	σ(volatility of stock)=
9	T(expiration date)=
10	t(excise date)=
11	D(Dividend)=
12	d1(non-dividend-paying)=
13	d2(non-dividend-paying)=
14	d1*(critical ex-dividend stock price)=
15	d2*(critical ex-dividend stock price)=
16	d1(dividend-paying)=
17	d2(dividend-paying)=
18	a1=
19	a2=
20	b1=
21	b2=
22	
23	c(St*,T-t;X)=
24	$c(S_t^*,T-t;X)-S_t^*-D+X=$
25	
26	$N_1(a_1)=$
27	$N_1(a_2)=$
28	$N_1(b_1)=$
29	$N_1(b_2)=$
30	$N_1(-b_1)=$
31	$N_1(-b_2)=$
32	ρ=
33	$a=a_1;b=-b_1$
34	$\phi(a,-b;-\rho)=$
35	$\phi(-a,b;-\rho)=$
36	$\rho_{ab}=$
37	$\rho_{ba}=$
38	$N_2(a,0;\rho_{ab})=$
39	$N_2(b,0;\rho_{ba})=$
40	δ=
41	$a=a_2;b=-b_2$
42	$\phi(a,-b;-\rho)=$
43	$\phi(-a,b;-\rho)=$
44	$\rho_{ab}=$
45	$\rho_{ba}=$
46	$N_2(a,0;\rho_{ab})=$
47	$N_2(b,0;\rho_{ba})=$
48	δ=
49	
50	$N_2(a_1,-b_1;\rho)=$
51	$N_2(a_2,-b_2;\rho)=$
52	
53	c(value of European call option to buy one share)=
54	p(value of European put option to sell one share)=
55	C(value of American call option to buy one share)=

Fig. 13.13 (continued)

	B
1	
2	
3	50
4	44.7557137976518
5	=B3-B11*EXP(-B7*B10)
6	45
7	0.08
8	0.2
9	=90/365
10	=50/365
11	1.5
12	=(LN(B3/B6)+(B7+0.5*B8^2)*B9)/(B8*SQRT(B9))
13	=B12-B8*SQRT(B9)
14	=(LN(B4/B6)+(B7+0.5*B8^2)*(B9-B10))/(B8*SQRT(B9-B10))
15	=B14-B8*SQRT(B9-B10)
16	=(LN(B5/B6)+(B7+0.5*B8^2)*B9)/(B8*SQRT(B9))
17	=B16-B8*SQRT(B9)
18	=(LN((B3-B11*EXP(-B7*B10))/B6)+(B7+0.5*B8^2)*(B9))/(B8*SQRT(B9))
19	=B18-B8*SQRT(B9)
20	=(LN((B3-B11*EXP(-B7*B10))/44.756)+(B7+0.5*B8^2)*(B10))/(B8*SQRT(B10))
21	=B20-B8*SQRT(B10)
22	
23	=B4*NORMSDIST(B14)-B6*EXP(-B7*(B9-B10))*NORMSDIST(B15)
24	=B23-B4-B11+B6
25	
26	=NORMSDIST(B18)
27	=NORMSDIST(B19)
28	=NORMSDIST(B20)
29	=NORMSDIST(B21)
30	=NORMSDIST(-B20)
31	=NORMSDIST(-B21)
32	=-SQRT(B10/B9)
33	
34	=Phi(-B20,0,-B37)
35	=Phi(-B18,0,-B36)
36	=((B32 * B18 - (-B20)) * IF(B18 >= 0, 1, -1))/ SQRT(B18 ^ 2 - 2 * B32 *B18 * -B20 +(-B20) ^ 2)
37	=((B32 * -B20 - (B18)) * IF(-B20 >= 0, 1, -1))/ SQRT(B18 ^ 2 - 2 * B32 *B18 * -B20 +(-B20) ^ 2)
38	=Bivarncdf(B18,0,B36)
39	=Bivarncdf(-B20,0,B37)
40	=(1 - IF(B18 >= 0, 1, -1) * IF(-B20 >= 0, 1, -1)) / 4
41	
42	=Phi(-B21,0,-B45)
43	=Phi(-B19,0,-B44)
44	=((B32 * B19 - (-B21)) * IF(B19 >= 0, 1, -1))/ SQRT(B19 ^ 2 - 2 * B32 *B19 * -B21 +(-B21) ^ 2)
45	=((B32 * -B21 - (B19)) * IF(-B21 >= 0, 1, -1))/ SQRT(B19 ^ 2 - 2 * B32 *B19 * -B21 +(-B21) ^ 2)
46	=Bivarncdf(B19,0,B44)
47	=Bivarncdf(-B21,0,B45)
48	=(1 - IF(B19 >= 0, 1, -1) * IF(-B21 >= 0, 1, -1)) / 4
49	
50	=Bivarncdf(B18,-B20,B32)
51	=Bivarncdf(B19,-B21,B32)
52	
53	=B5*NORMSDIST(B16)-B6*EXP(-B7*B9)*NORMSDIST(B17)
54	=B6*EXP(-B7*B9)*NORMSDIST(-B17)-B5*NORMSDIST(-B16)
55	=(B3-B11*EXP(-B7*B10))*(NORMSDIST(B20)+Bivarncdf(B18,B20,-SQRT(B10/B9)))-B6*EXP(-B7*B9)*(NORMSDIST(B21)*EXP(B7*(B9-B10))+Bivarncdf(B19,-B21,-SQRT(B10/B9)))+B11*EXP(-B7*B10)*NORMSDIST(B21)

Fig. 13.13 Microsoft Excel program for calculating two alternative American call options

Chapter 14
Simple Linear Regression and Correlation: Analyses and Applications

Chapter Outline

Key Terms

Unbiased estimate	Confidence interval
Standard deviation of error terms	Prediction interval
Standard errors of estimate	Market model
Prediction	Mean response
Forecast	Individual response
Conditional prediction	Measurement errors
Forecast error	Proxy errors
Confidence belt	

14.1 Introduction

This chapter clarifies and expands on the material presented in Chap. 13 by providing calculations, analyses, and applications to business and economics.

First, statistics used to test the significance of the intercept (a), slope (b), and simple correlation coefficient (r) are derived, and the use of these statistics is

C.-F. Lee et al., *Statistics for Business and Financial Economics*, 675
DOI 10.1007/978-1-4614-5897-5_14, © Springer Science+Business Media New York 2013

demonstrated. Second, confidence intervals for alternative prediction methods in terms of simple regression are investigated. Third, applications of simple regression in business are explored in some detail. Finally, an example is offered to show how the statistical computer programs MINITAB and SAS can be used to do simple regression analyses.

14.2 Tests of the Significance of α and β

Chapter 13 detailed basic concepts and estimation procedures for the linear regression line, regression parameters, and correlation coefficients. Now we will discuss statistical tests involving regression parameters, intercept (a), and slope (b). (The sample regression coefficients a and b are estimates of population regression coefficients α and β, just as $\bar{x}$ is the estimate of μ.) In addition, these coefficients have sampling distributions (just as $\bar{x}$ has a sampling distribution). With the usual regression assumptions, the sampling distributions of a and b have the following properties:

1. a and b are *unbiased estimates* of α and β, that is,[1] $E(a) = \alpha$ and $E(b) = \beta$. This means that the expected value of a; is equal to the population parameter α and that the expected value of b is β. On average, then, the value obtained from the estimators is equal to the population parameters. Some estimates will be too low and some will be too high, but there is no systematic bias.
2. The sampling distributions of a and b are normally distributed with means α and β and variances S_a^2 and S_b^2:

$$S_a^2 = S_e^2 \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{14.1}$$

$$S_b^2 = \frac{S_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{14.2}$$

where s_e is the estimate of the *standard deviation of error terms* for a regression, as defined in Eq. 13.20. s_a and s_b are *standard errors of estimate* for a and b. The sampling distribution of b is presented in Fig. 14.1.

[1] If there are measurement errors or proxy errors associated with the independent variable, then b is no longer an unbiased estimate for β. See Appendix 1.

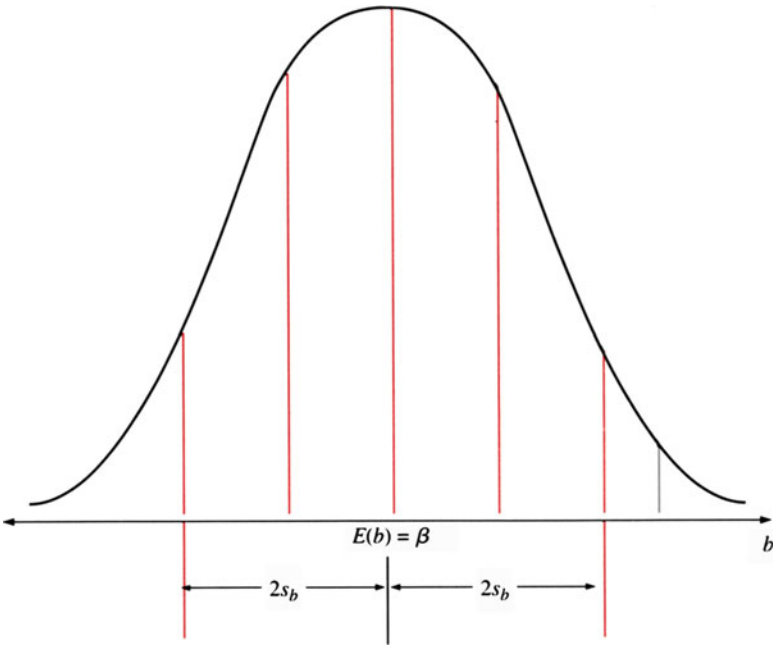


Fig. 14.1 Sampling distribution of b

14.2.1 Hypothesis Testing and Confidence Interval for β and α

The slope β is the parameter of most interest to business, economic, and other statisticians, because it can be used to measure the relationship between dependent and independent variables. The slope β measures the change in y that results from a 1-unit change in x. If β is equal to $\frac{1}{2}$, then $E(\mathrm{Y} \mid \mathrm{X})$ changes by plus $\frac{1}{2}$ unit when x is increased by 1 unit.

For example, let the population regression function be

$$E(Y|X = x) = \mu_{Y|x} = \alpha + \beta x$$

where $\mu_{Y|x}$ is the population mean of Y, given x. (Once again, assume x is height and Y *is* weight.) Then β shows the increase in weight when there is a unit increase (a 1-in. increase) in height. As another example, let Y be consumption and x be income. Then β shows the increase in consumption when there is a unit increase in income. If β is equal to 0.75, then consumption is expected to go up 75 cents when income increases by 1 dollar.

In business and economic research, we need a guideline to help us determine whether the independent variable X is useful in predicting the value of Y. Suppose the population relationship is such that $\beta = 0$. This means the population regression

Table 14.1 Weight and height data

Height, x_i (in.)	Weight, y_i (pounds)
55	92
56	95
57	99
58	97
59	102
60	104

line must be horizontal, that is, $\hat{Y} = \bar{Y}$. When $\beta = 0$, the value of X is of no help in predicting Y: no matter how much X changes, there is no change in Y (on the average). Thus, determining whether $\beta = 0$ often proves beneficial, but how is such a determination made?

14.2.1.1 Two-Tailed z-Test Versus Two-Tailed *t*-Test for β

When researchers want to test whether the slope is different from zero, the alternative hypothesis (H_1) is that the slope is different from zero; it does not matter whether the slope is positive or negative. The null hypothesis (H_0) in such cases is that the slope is zero.

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

If b is normally distributed with $E(b) = \beta$ and variance s_b^2 (these were described earlier in this section as the two properties of b), we can graph the sampling distribution as shown in Fig. 14.1. We may use either the z statistic or the t statistic to test whether the null hypothesis (H_0: $\beta = 0$) is true.

To perform this test, we will again use the sample data of heights and weights from Chap. 13, restated now in Table 14.1.

The worksheet for calculating a and b for the data of Table 14.1 is given in Table 14.2. From Table 14.2, we can obtain

$$s_x^2 = \frac{17.5}{5} = 3.5$$
$$s_y^2 = \frac{98.8334}{5} = 19.7667$$
$$s_{xy} = \frac{39.5}{5} = 7.9$$
$$r = \frac{7.9}{\sqrt{(3.5)(19.7667)}} = .9498$$
$$b = \frac{s_{xy}}{s_x^2} = \frac{7.9}{3.5} = 2.2571$$

Table 14.2 Worksheet for calculation of the coefficients a and b

y_i	x_i	$(y_i - \bar{y})$	$(x_i - \bar{x})$	x_i^2	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
92	55	−6.1667	−2.5	3,025	38.0282	6.25	15.4168
95	56	−3.1667	−1.5	3,136	10.0280	2.25	4.7501
99	57	0.8333	−0.5	3,249	0.6944	0.25	–0.4167
97	58	−1.1667	0.5	3,364	1.3612	0.25	–0.5834
102	59	3.8333	1.5	3,481	14.6942	2.25	5.7499
104	60	5.8333	2.5	3,600	34.0274	6.25	14.5833
Sum 589	345	–0.0002	0	19,855	98.8334	17.50	39.50
Mean 98.1667	57.5	–	–	–	–	–	–

The sample slope estimate (b) equals 2.2571 lb/in. To test $E(b) = \beta = 0$, we follow the hypothesis-testing technique discussed in Chap. 11. If the sample size is large, we can find the z statistic.

$$z = [b - E(b)]/S_b = (b - 0)/S_b = b/S_b \tag{14.3}$$

From Tables 13.6 and 14.2, we obtain $s_e^2 = 9.67618/(6 - 2) = 2.4191$ and $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 17.5$. Substituting these estimations into Eq. 14.2 yields

$$s_b = \sqrt{\frac{2.4191}{17.5}} = .3718$$

Then the z statistic becomes 2.2571/.3718 = 6.0707, which shows that b = 2.2571 lb is 6.0707 standard deviations away from $E(b) = \beta = 0$. Under the significance level $\alpha = .05$, from Table 3 of Appendix A, we have $Z_{.05} = 1.96$. Because $6.0707 > 1.96$, we conclude that it is highly improbable that $b = 2.2571$ lb/in. came from a population with $\beta = 0$ and we reject H_0. That is, we accept H_1, which is $\beta \neq 0$, and conclude that the independent variable *is* useful in predicting the dependent variable.

It should be noted that when the sample size is as small as in this example, we should use a t statistic instead of a z statistic because we are using the sample estimate of σ_e – that is, s_e – to calculate s_b. The value of s_b is a measure of the amount of sampling error in the regression coefficient b, just as s_x was a measure of the sampling error of $\bar{x}$. By using the t-test, we can redefine Eq. 14.3 as

$$t_{n-2} = \frac{b - 0}{s_b} \tag{14.4}$$

The t statistic, t_{n-2}, which follows a t distribution with $(n-2)$ degrees of freedom, was discussed in Chap. 9. If the sample size is large, then the difference between the t statistic and the z statistic indicated in Eq. 14.3 is small enough for us to use the z statistic in testing the null hypothesis.

The t statistic associated with β in terms of Eq. 14.4 is

$$t_{n-2} = \frac{2.2571}{.3718} = 6.0707$$

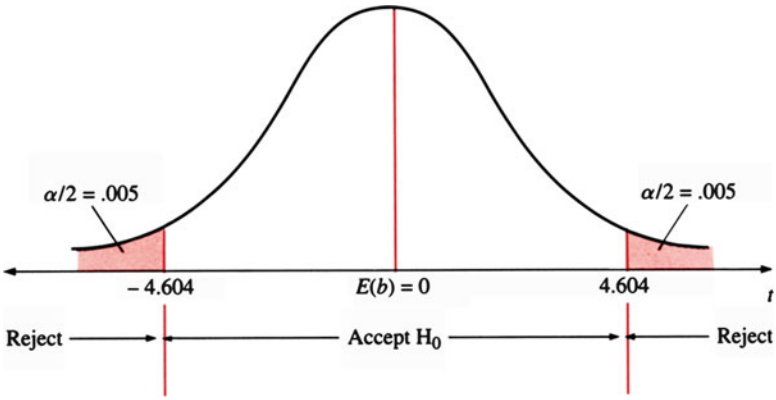


Fig. 14.2 Two-tailed test of estimated slope (b)

Using this information, we can perform a two-tailed t-test as follows:

$$H_0 : \beta = 0 \text{ versus } H_1 : \beta \neq 0$$

Before we perform the null hypothesis test, however, we must specify the significance level α. We choose 0.01. Because a two-tailed test is used, the regression region on the right tail and left tail has an area of .01/2 or 0.005. When the degrees of freedom is $v = (n-2) = 4$, then $t_{.01/2,4} = t_{.005,4} = 4.604$. The estimated t_{n-2} is 6.0707, which is larger than the absolute value of both $-t_{.005,4}$ and $t_{.005,4}$, as indicated in Fig. 14.2. Hence, we can conclude that the estimated slope is significantly different from zero when α is equal to 1 % under the two-tailed test. In other words, the regression line does improve our ability to estimate the dependent variable, weight.

14.2.1.2 Two-Tailed t-Test for β

Similarly to the null hypothesis test for β, we can define the hypotheses in a t-test for α as

$$H_0 : \alpha = 0 \text{ versus } H_1 : \alpha \neq 0$$

A t statistic to test whether the population intercept, α, is significantly different from zero is

$$t_{n-2} = \frac{a - 0}{s_a}, \tag{14.5}$$

where s_a is the standard deviation of the sample intercept a. Using the data of Table 14.2, we can estimate the intercept a and its standard error s_a as follows:

$$a = \bar{y} - b\bar{x} = 98.1667 - (2.2571)(57.5) = -31.6166$$

$$s_a = \sqrt{\frac{(2.4191)(19,855)}{(6)(17.5)}} = 21.3879$$

To test whether the intercept, α, is equal to zero, we divide $s_a = 21.3879$ into $a = -31.6166$, obtaining

$$t = \frac{-31.6166}{21.3879} = -1.4782$$

Because -1.4782 is larger than $-2.776 = t_{.05/2,4} = t_{.025,4}$ (from Table A4 in Appendix A at the end of this book), we conclude that the estimated intercept, a, is not significantly different from zero under a two-tailed t-test with $\alpha = .05$.

14.2.1.3 One-Tailed *t*-Test for β

A researcher sometimes uses a one-tailed hypothesis test where the alternative test is that the slope is greater than zero or less than zero.

$$H_0 : \beta = 0 \text{ versus } H_1 : \beta > 0 \text{ or } \beta < 0$$

For $(n-2) = 6-2 = 4$ degrees of freedom, the probability that t is larger than 6.0707 falls below .005 (see Table 4 in Appendix A). Thus, it is highly unlikely that $b = 2.2571$ will occur by chance when $\beta = 0$ and we can reject H_0 and accept H_1.

For the one-tailed test, the critical values are $t_{.005,4} = 4.604$ and $t_{.01,4} = 3.747$. Again, $t_4 = 6.0707$ is larger than both 3.747 and 4.604. Hence, we can also conclude that the estimated slope is significantly different from zero when $\alpha = .5$ % or $\alpha = 1$ % under a one-tailed test. Incidentally, using the nonnegative one-tailed test makes more sense, because it is not reasonable to expect an inverse relationship between height and weight.

Figure 14.3 shows the critical t-value for $\alpha = .005$ with 4 degrees of freedom. Our estimated t-value is equal to 6.0707, and it is larger than 4.604, so it falls within the rejection region when $\alpha = .5$ %.

14.2.1.4 Confidence Interval for β

On the basis of the confidence interval concepts discussed in Chap. 10, we obtain the confidence interval for b as

$$b - t(\alpha/2, n-2)s_b \leq \beta \leq b + t(\alpha/2, n-2)s_b \tag{14.6}$$

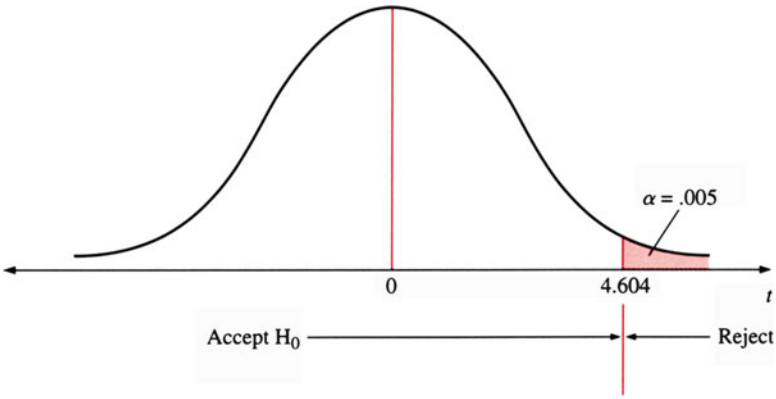


Fig. 14.3 One-tailed test for regression slope

A 95 % confidence interval of a two-tailed test, given that $n = 6$, $t(.025, 4) = 2.776$ and $s_b = .3718$, is

$$2.2571 - (2.776)(.3718) \leq \beta \leq 2.2571 + (2.776)(.3718)$$
$$1.2250 \leq \beta \leq 3.2892$$

Thus, an increase in weight of between 1.2250 and 3.2892 lb for each 1-in. increase in height can be expected. It should be noted that this result is based on only 6 observations and that, all other things being equal, precision would be greater if n were larger.

Similarly, the 99 % confidence interval of a two-tailed test, given $n = 6$, $t_{0.005.4} = 4.604$, and $s_b = .3718$, is

$$2.2571 - (4.604)(.3718) \leq \beta \leq 2.2571 + (4.604)(.3718)$$
$$.5433 \leq \beta \leq 3.9689$$

Figure 14.4 shows only the 99 % confidence intervals for the population regression slope, calculated from the height and weight data in Table 14.1. Note that the 99 % confidence interval is wider than the 95 % confidence interval.

14.2.2 The F*-Test Versus the* t*-Test*

Besides using the t statistic to test whether regression slope is significantly different from zero, we can also use the F statistic to test whether the regression slope is significantly different from zero. In this section we will discuss how the F statistic is used to perform the test and how the F-test is related to the t-test.

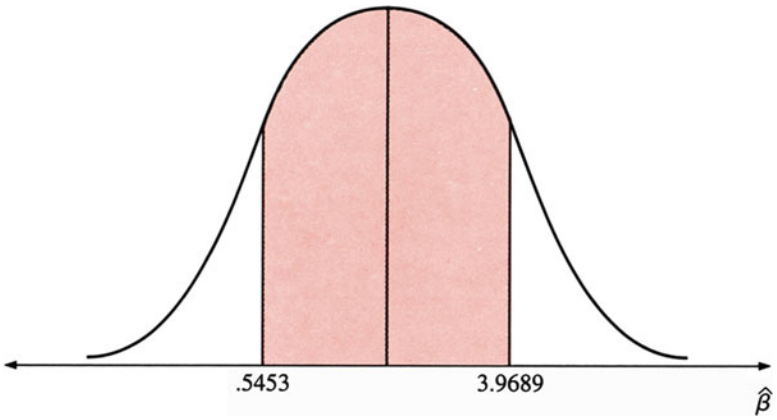


Fig. 14.4 Confidence interval for regression slope

Table 14.3 Analyses of variance

Actual y_i	Estimated $\hat{y}_i$	$(\hat{y}_i - y_i)^2$	$(y_i - \bar{y})^2$
92	92.5239	0.27447	38.0282
95	94.7810	0.04796	10.0280
99	97.0381	3.84905	0.6944
97	99.2952	5.26794	1.3612
102	101.5523	0.20043	14.6942
104	103.8094	0.03633	34.0274
Total		9.67618	98.8334

Sources of variation	Sum of squares	Degrees of freedom	Mean square
Regression	$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k	SSR/k
Residual	$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n - k - 1$	SSE/$(n - k - 1)$
Total	$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n - 1$	SST/$(n - 1)$
Regression	89.1572	1	89.1572
Residual	9.6762	4	2.4191
Total	98.8334	5	19.7667

14.2.2.1 Procedure for Using the *F*-Test

From Eq. 13.17 and Table 13.6 in the last chapter, we have

$$\begin{array}{ccccc} \text{SST} & = & \text{SSE} & + & \text{SSR} \\ \text{Total} & & \text{Unexplained} & & \text{Explained} \\ \text{Variation} & & \text{Variation} & & \text{Variation} \end{array} \tag{14.7}$$

Recall that the degrees of freedom associated with SST and SSE are $(n-1)$ and $(n-2)$, respectively. For convenience, we repeat Table 13.6 here as Table 14.3. From Table 14.3, we know that the degrees of freedom for SST must equal the sum of SSE and SSR; therefore, the number of degrees of freedom for SSR equals 1 (the

number of independent variables). As we noted in Chaps. 12 and 13, a sum of squares divided by its degrees of freedom is called a mean square. There are three different mean squares for a regression analysis, as indicated in Table 14.3.

Table 14.3 shows that an ANOVA table can be constructed for both regression analysis and analysis of variance. If the estimated slope b is not significantly different from zero, then

$$\begin{aligned}\hat{y}_i &= a + bx_i \\ &= \bar{y} + b(x_i - \bar{x}) \\ &= \bar{y}\end{aligned} \tag{14.8}$$

This implies that $\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ is small and that SST approaches SSE. Therefore, we can use the ratio F of Eq. 14.9 to test whether the estimated slope b is significantly different from zero:

$$F_{(1,n-2)} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{\text{MSR}}{\text{MSE}} \tag{14.9}$$

where

MSR = mean square regression (due to regression)
MSE = s_e^2 = mean square error (residuals)

From Chaps. 9, 12, and 13, we know that the statistic $F_{(1,n-2)}$ is an F distribution with 1 and $(n-2)$ degrees of freedom. Using the empirical results of Table 14.3, we calculate the F-value as

$$F = \frac{89.1572/1}{9.6762/4} = 36.8563$$

To use this estimated F-value to test whether $b = 2.2571$ is significantly different from zero, we first choose a significance level of α and then use the F table, Table A6 of Appendix A, to determine the critical value. If we choose a significance level of $\alpha = .01$, the critical value is $F_{1,4} = 21.2$.

As shown in Fig. 14.5, the decision rule is to reject H_0 – that is, to accept the hypothesis that the regression line does contribute to an explanation of the variation of y – if the calculated value of F based on our height and weight example exceeds 21.2 (see Fig. 14.5). Because the sample value of F, 36.8563, is larger than the critical value, 21.2, we reject the null hypothesis and conclude that the height does indeed help explain the variation of weight.

14.2.2.2 The Relationship Between the *F*-Test and the *t*-Test

The F-test on the variation ratio between mean square regression (MSR) and mean square error (MSE), as defined in Eq. 14.9, is comparable to the t-test on the

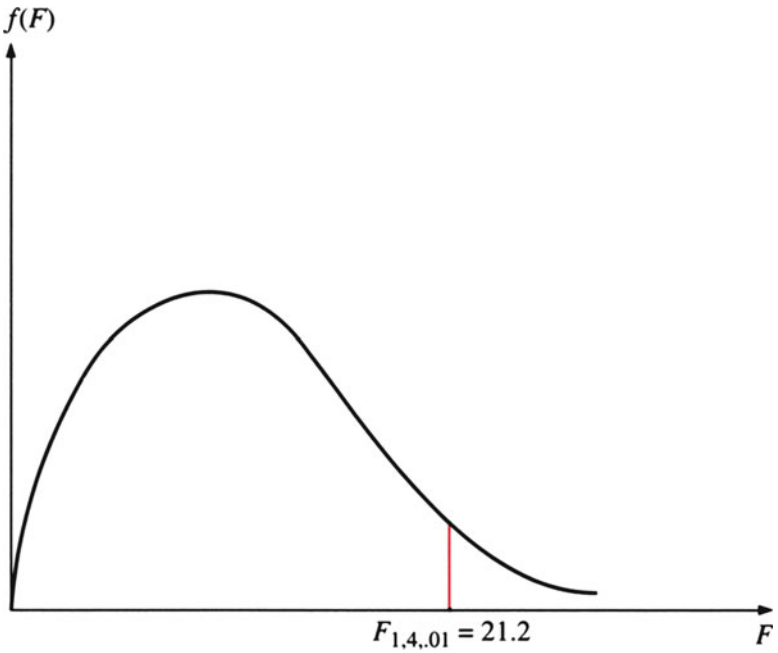


Fig. 14.5 Critical value for F-test

significance of the slope: both test whether the slope is significantly different from zero. Actually, the t-test and the F-test on the variance ratio between MSR and MSE are equivalent tests for the significance of the linear relationship between two variables x and y. This will be explained further in Appendix 2.

The advantage of using the F-test is that it can be generalized to test a set of regression coefficients associated with multiple regression, which is discussed in the next chapter. In addition, the F-test can also be used to investigate other important topics in regression analysis (Chap. 16). However, the t-test, rather than the F-test, is used to test whether β differs from a specific value other than zero.

14.3 Test of the Significance of ρ

So far in this chapter, we have used the z statistic and the t statistic to test H_0: $\beta = 0$. There is also a t-test and a z-test to test H_0: $\rho = 0$. In other words, these tests are used to determine whether ρ, the population correlation coefficient between two variables, is statistically significantly different from zero.

14.3.1 t-Test for Testing $\rho = 0$

If x and y are bivariate normally distributed, then we can use the t-distributed random variable t_{n-2} to test whether ρ is statistically significantly different from zero.

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{14.10}$$

where r is the sample correlation coefficient between x and y and n represents the number of observations. Using our example of children's height and weight, $r = .9498$ and $n = 6$, we find that

$$t_4 = \frac{.9498\sqrt{6-2}}{\sqrt{1-(.9498)^2}} = \frac{1.8996}{.3129} = 6.0709$$

Table A4 of Appendix A gives the critical values for 4 degrees of freedom, at $\alpha = .05$ and $\alpha = .025$, as

$$t_{4,.05} = 2.132 \text{ and } t_{4,.025} = 2.776$$

Therefore, the null hypothesis of no relationship between x and y can be rejected against the alternative that the true correlation is positive at both 5 % and 2.5 % significance levels. The height and weight data, then, contain fairly strong evidence supporting the hypothesis of a positive (linear) association between students' heights and weights.

The t-value for testing H_0: $\rho = 0$ is equal to that used for testing H_0: $\beta = 0$ unless there are rounding errors. That is, $t_{n-2} = b/s_b$.[2]

[2] From Eq. 13.26, we have

$$r = b\left(s_x/s_y\right) \text{ and } r^2 = b^2\left(s_x^2/s_y^2\right)$$

Substituting these two equations into Eq. 14.10 and rearranging the terms, we obtain

$$t_{n-2}\frac{b}{\sqrt{\frac{s_y^2-b^2s_x^2}{(n-2)\left(s_x^2\right)}}} = \frac{b}{\sqrt{s_e^2\Big/\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{b}{s_b}$$

14.3.2 z-Test for Testing $\rho = 0$ or $\rho = \text{Constant}$

The t-value of vindicated in Eq. 14.10 cannot be used for making confidence statements about sample correlations. In addition, this approach is not suitable for testing a null hypothesis other than $\rho = 0$, such as $\rho = .20$ or $\rho_1 - \rho_2 = 0$. A convenient and sufficiently accurate solution of these problems was provided by Fisher (1921), who derived a transformation from r to a quantity, h, distributed almost normally with variance[3]

$$s_h^2 = \frac{1}{h-3} \tag{14.11}$$

where

$$h = \frac{1}{2}\left[\log_e\left(\frac{1+r}{1-r}\right)\right] \tag{14.12}$$

Note that the variance of h is independent of the value of the correlation in the population from which the sample is drawn (in contrast to the variance of r, which is dependent on the population). If $r = 0$, then $h = 0$; therefore, the null hypothesis of testing that $h = 0$ is identical to testing that $r = 0$. In our example,

$$r = .9498, h = \frac{1}{2}\log_e\frac{(1+.9498)}{(1-.9498)} = 1.8297$$

Using Eq. 14.11, we have

$$s_h = \sqrt{\frac{1}{6-3}} = .5773$$

Dividing .5773 into 1.8297, we obtain

$$z = 1.8297/.5773 = 3.1694$$

We are testing the hypothesis that $\rho = 0$ against the alternative that $\rho \neq 0$. From the t distribution table, Table A4 in Appendix A, we find the critical value is $t_{.005} = 2.576$, which corresponds to degrees of freedom (df) $= \infty$. Approximately this value can also be taken from the normal distribution table, Table A3 of Appendix A. These results imply that the correlation coefficient .9498 is significantly different from zero at the $\alpha = 1$ % significance level. Finally, the method employed in Eq. 14.6 can be used to obtain confidence intervals for h.

[3] Fisher R.A.: On the probable error of a correlation coefficient deduced from a small sample. Mentor **1**, part 4, 3–32 (1921)

A 99 % confidence interval for h, given that $h = 1.8297$, $z_{.005} = 2.576$, and $s_h = .5773$, is

$$1.8297 - (2.576)(.5773) \leq h \leq 1.8297 + (2.576)(.5773)$$

or

$$.3426 \leq \mathrm{h} \leq 3.3168$$

Because this interval does not include $h = 0$, it also does not include $r = 0$. Again, this implies that the correlation coefficient .9498 is significantly different from zero at $\alpha = 1$ %.

14.4 Confidence Interval for the Mean Response and Prediction Interval for the Individual Response

In this section, we discuss both point estimates and confidence intervals for the mean response. We also consider point estimates and prediction intervals for the individual response.

14.4.1 Point Estimates of the Mean Response and the Individual Response

One of the important uses of a sample regression line is to obtain *predictions* (or *forecasts)* for the dependent variable, conditional on an assumed value of the independent variable. This kind of prediction is called a *conditional prediction* (or conditional forecast). Suppose the independent variable is equal to some specified value x_{n+1} and that the linear relationship between y_t and x_t continues to hold.[4] Then the corresponding value of the dependent variable Y_{n+1} is

$$Y_{n+1} = \alpha + \beta x_{n+1} + \varepsilon_{n+1} \tag{14.13}$$

which, given x_{n+1}, has the conditional expectation

$$E(Y_{n+1}|x_{n+1}) = \alpha + \beta x_{n+1} \tag{14.14}$$

[4] x_{n+1} can be a given value or forecasted value. If a regression is used to describe a time-series relationship, then x_{n+1} is a forecasted value. This issue will be discussed in detail later in this chapter.

Equation 14.14 can be used to estimate the conditional expectation $E(Y_{n+1}|x_{n+1})$ when the independent variable is fixed at x_{n+1}; Eq. 14.13 can be used to estimate the actual value for a given independent variable x_{n+1}. In other words, Eq. 14.14 is used to estimate the *mean response* and Eq. 14.13 to estimate the *individual response*. For both problems, we can obtain both the point estimate and the interval estimate.

To obtain the best point estimate, we should first estimate the sample regression line

$$y_i = a + bx_i + e_i \tag{14.15}$$

Then we can substitute the given value x_{n+1} into the estimated Eq. 14.15 and obtain

$$\hat{y}_{n+1} = a + bx_{n+1} \tag{14.16}$$

This is the best point estimate for forecasts of both conditional expectation (mean response) and actual value (individual response). The forecast of conditional expectation value is equal to the forecast of actual expectation value. However, they are interpreted differently. This different interpretation will become important when we investigate the process of making interval estimates.

14.4.2 Interval Estimates of Forecasts under Three Cases of Estimated Variance

To construct a confidence interval for forecasts, it is necessary to know the distribution, the mean, and the variance of $\hat{y}_{n+1}$. The distribution of $\hat{y}_{n+1}$ is a *normal* distribution. The variance associated with $\hat{y}_{n+1}$ can be classified into three cases. Let's examine them individually.

Case 14.1 Conditional Expectation (Mean Response) with $x_{n+1} = \bar{x}$

From the definitions of the intercept of a regression and the sample regression line, we have

$$\hat{y}_{n+1} = \bar{y} - b\bar{x} + bx_{n+1} = \bar{y} + b(x_{n+1} - \bar{x})$$

If $x_{n+1} = \bar{x}$, then we have $\hat{y}_{n+1} = \bar{y}$. Following Appendix 3, we obtain the estimate of the variance for y_{n+1} as

$$s^2(\hat{y}_{n+1}) = s^2(\bar{y}) = s_e^2/n \tag{14.17}$$

Case 14.2 Conditional Expectation (Mean Response) with $x_{n+1} = x$

In this case, the forecast value can be defined as

$$\hat{y}_{n+1} = \bar{y} + b(x_{n+1} - \bar{x})$$

Following Appendix 3, we obtain the estimate of the variance for y_{n+1} as

$$\begin{aligned} s^2(\hat{y}_{n+1}) &= s^2[\bar{y} + b(x_{n+1} - \bar{x})] \\ &= s_e^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \end{aligned} \tag{14.18}$$

This is the variance for the estimated dependent variable $(\hat{y}_{n+1})$.

Case 14.3 Actual Value of $y_{n+1,i}$ (Individual Response)

In this case, we want to predict the actual value of y_{n+1}. The procedure for finding the variance is to find the variance of the difference $\hat{y}_{n+1} - y_{n+1,i}$ – in other words, the *forecast error*. The sample variance of residual $(\hat{y}_{n+1} - y_{n+1,i})$ can be defined as

$$\begin{aligned} s^2(\hat{y}_{n+1,i}) = s^2(\hat{y}_{n+1} - y_{n+1,i}) &= s_e^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) + s_e^2 \\ &= s_e^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \end{aligned} \tag{14.19}$$

Using Eqs. 14.17, 14.18, and 14.19, we can obtain a confidence interval for the mean response and a prediction interval for the individual response as follows. For case 1 of the mean response, the confidence interval is

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \frac{s_e}{\sqrt{n}} \tag{14.20}$$

For case 2 of the mean response, the confidence interval is

(continued)

Case 14.3 (continued)

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{14.21}$$

For case 3, the individual response of actual value $y_{n+1,i}$, and the prediction interval is

$$\hat{y}_{n+1,i} \pm t_{n-2,\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{14.22}$$

To show how Eqs. 14.20, 14.21, and 14.22 can be applied, we will now use the height and weight example to estimate the variances and the prediction intervals.

14.4.3 Calculating Standard Errors

Table 14.4 is the worksheet for 3 alternative forecasts that we generate by using the data of Table 14.1. From Table 14.2,

$$\bar{x} = 57.5, \quad \sum_{i=1}^{n} (x_i - \bar{x})^2 = 17.5$$

Let $x_{n+1} = 55, 56, 57, 58, 59$, and 60. Substituting these data into Eq. 14.16, we obtain $\hat{y}_{n+1}$ as indicated in column (3) of Table 14.4. Substituting $s_e^2 = 2.4191$ and $n = 6$ into Eq. 14.17, we obtain $s(\bar{y})$ as indicated in column (5) of Table 14.4.

To calculate $s^2(\hat{y}_{n+1}|x_{n+1} \neq \bar{x})$ and $s^2(\hat{y}_{n+1,i})$, we substitute related information into Eqs. 14.21 and 14.22 as follows:

1. $x_{n+1} = 55$

$$s^2(\hat{y}_{n+1}) = s_e^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$= (2.4191) \left(\frac{1}{6} + \frac{(55 - 57.5)^2}{17.5} \right) = 1.2671$$

Table 14.4 Worksheet for calculating three alternative standard errors

(1) x	(2) y	(3) $\hat{y}_{n+1}$	(4) s_e^2	(5) $s(\bar{y})$	(6) $s(y_{n+1})$	(7) $s(\hat{y}_{n+1,i})$
55	92	92.5239	2.4191	0.6350	1.1257	1.9200
56	95	94.7810	2.4191	0.6350	0.8452	1.7702
57	99	97.0381	2.4191	0.6350	0.6617	1.6903
58	97	99.2952	2.4191	0.6350	0.6617	1.6903
59	102	101.5523	2.4191	0.6350	0.8452	1.7702
60	104	103.8094	2.4191	0.6350	1.1257	1.9200

$$s^2(\hat{y}_{n+1,i}) = s_e^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$
$$= 1.2671 + 2.4191 = 3.6862$$

2. $x_{n+1} = 56$

$$s^2(\hat{y}_{n+1}) = (2.4191)\left(\frac{1}{6} + \frac{(56 - 57.5)^2}{17.5}\right) = .7144$$
$$s^2(\hat{y}_{n+1,i}) = 2.4191 + .7144 = 3.1335$$

3. $x_{n+1} = 57$

$$s^2(\hat{y}_{n+1}) = (2.4191)\left(\frac{1}{6} + \frac{(57 - 57.5)^2}{17.5}\right) = .4379$$
$$s^2(\hat{y}_{n+1,i}) = 2.4191 + .4379 = 2.8570$$

4. $x_{n+1} = 58$

$$s^2(\hat{y}_{n+1}) = (2.4191)\left(\frac{1}{6} + \frac{(58 - 57.5)^2}{17.5}\right) = .4379$$
$$s^2(\hat{y}_{n+1,i}) = 2.4191 + .4379 = 2.8570$$

5. $x_{n+1} = 59$

$$s^2(\hat{y}_{n+1}) = (2.4191)\left(\frac{1}{6} + \frac{(59 - 57.5)^2}{17.5}\right) = .7144$$
$$s^2(\hat{y}_{n+1,i}) = 2.4191 + .7144 = 3.1335$$

6. $x_{n+1} = 60$

$$s^2(\hat{y}_{n+1}) = (2.4191)\left(\frac{1}{6} + \frac{(60 - 57.5)^2}{17.5}\right) = 1.2671$$

$$s^2(\hat{y}_{n+1,i}) = 2.4191 + 1.2671 = 3.6862$$

Alternative estimates of $s(\hat{y}_{n+1})$ and $s(\hat{y}_{n+1,i})$ are listed in columns (6) and (7) of Table 14.4, respectively.

14.4.4 *Confidence Interval for the Mean Response and Prediction Interval for the Individual Response*

Let $\alpha = .05$, then $t_{.025,4} = 2.776$. Substituting all related information into the formulas for confidence interval and prediction interval as shown in Eqs. 14.20, 14.21, and 14.22, we can obtain 95 % confidence interval estimates for the mean response and individual response as shown in the cases that follow.

Case 14.4 The Mean Response with $x_{n+1} = \bar{x}$

Because $x_{n+1} = \bar{x} = 57.5$ and

$$\hat{y}_{n+1} = -31.6166 + (2.2571)(57.5) = 98.1667$$

we can use $s_e/\sqrt{n} = .6350$ as indicated in column (5) of Table 14.4 to define the 95 % confidence interval in terms of Eq. 14.20 as

$$98.1667 - (2.776)(.6350) < E(Y_{n+1}|\bar{x}) < 98.1667 + (2.776)(.6350)$$

$$96.4038 < E(Y_{n+1}|\bar{x}) < 99.9295$$

How do we interpret this interval? We say that if 100 random samples of size 6 are selected and the confidence intervals of Eq. 14.20 are constructed, we should expect 95 % of those intervals to contain $E(Y_{n+1}|\ \bar{x} = 57.5)$. The confidence interval calculated here is one of the 100 such intervals. This confidence interval is indicated in Fig. 14.6, where A and *B* represent the lower bound and upper bound, respectively.

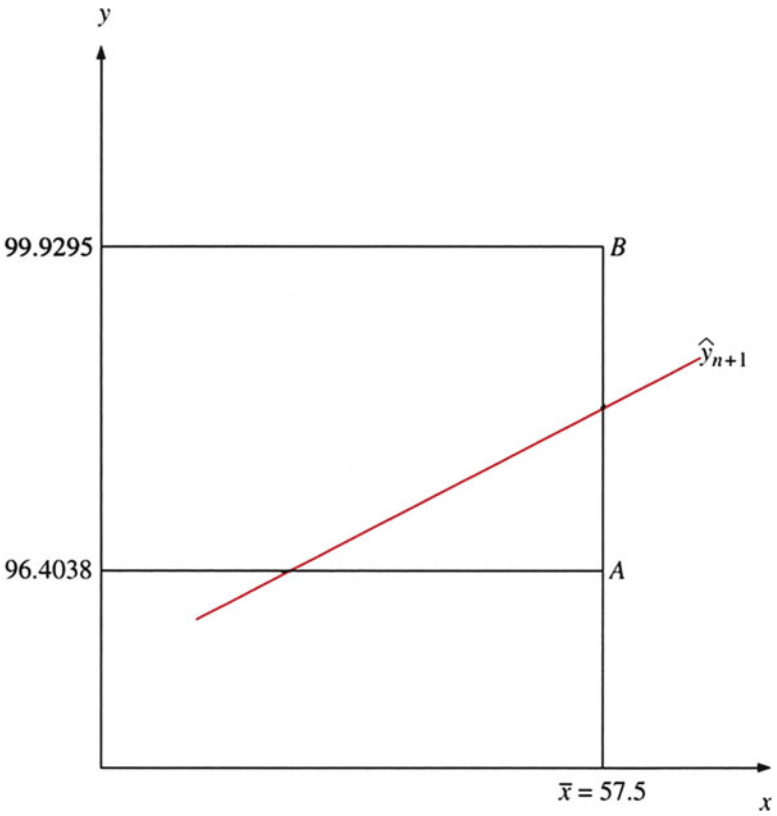


Fig. 14.6 Confidence interval for mean response with $x_{n+1} = \bar{x}$

Case 14.5 The Mean Response with $x_{n+1} = x$

In this case, the confidence intervals depend on the values of x and y_{n+1}. Using alternative standard errors as indicated in column (6) of Table 14.4, we find that the 95 % confidence intervals for this case are

$$x = 55:$$

$$92.5239 - (2.776)(1.1257) < E(Y_{n+1}|x_{n+1}) < 92.5239 + (2.776)(1.1257)$$

$$89.3990 < E(Y_{n+1}|x_{n+1}) < 95.6488$$

$$x = 56: \quad 92.4347 < E(Y_{n+1}|x_{n+1}) < 97.1273$$

$$x = 57: \quad 95.2012 < E(Y_{n+1}|x_{n+1}) < 98.8750$$

(continued)

Case 14.5 (continued)

$$x = 58:\ 97.4583 < E(Y_{n+1}|x_{n+1}) < 101.1321$$

$$x = 59:\ 99.2060 < E(Y_{n+1}|x_{n+1}) < 103.8986$$

$$x = 60:\ 100.6845 < E(Y_{n+1}|x_{n+1}) < 106.9343$$

How are these intervals interpreted? If 100 samples of size 6 are selected and 6 confidence intervals of Eq. 14.21 corresponding to $x = 55, 56, \ldots, 60$ are constructed, we should expect 95 of them to contain $E(Y_{n+1} \mid x_{n+1})$ for a given x. Each interval estimate here is one of the 100 such intervals for a given x. These confidence intervals are graphically presented in Fig. 14.7.

When we connect the points, we get a *confidence belt* that is symmetric in width around the value $X_{n+1} = \bar{x}$. Note that this confidence belt was constructed in a single sample. Each time a new sample is selected, there is a new confidence belt.

Case 14.6 The Individual Response

The standard error here is larger than that of Case 14.5 by an amount s_e^2. Using the alternative standard errors indicated in column (7) of Table 14.4, we find that the 95 % prediction intervals for this case are

$$x = 55:$$

$$92.5239 - (2.776)(1.1257) < E(Y_{n+1}|x_{n+1}) < 92.5239 + (2.776)(1.9200)$$

$$87.1940 < E(Y_{n+1}|x_{n+1}) < 97.8538$$

$$x = 56:\ 89.8669 < E(Y_{n+1}|x_{n+1}) < 99.6951$$

$$x = 57:\ 92.3458 < E(Y_{n+1}|x_{n+1}) < 101.7304$$

$$x = 58:\ 94.6029 < E(Y_{n+1}|x_{n+1}) < 103.9875$$

$$x = 59:\ 96.6382 < E(Y_{n+1}|x_{n+1}) < 106.4664$$

$$x = 60:\ 98.4795 < E(Y_{n+1}|x_{n+1}) < 109.1393$$

(continued)

Case 14.6 (continued)

The interpretations of these prediction intervals are similar to those of the confidence intervals of Case 14.5, but these intervals are wider. They are shown in Fig. 14.8.

As in Fig. 14.7, we can construct a confidence belt by connecting the points. The confidence belt of Case 14.6 is for the forecasts of the actual values of the students' weights. The confidence belt of Case 14.5 is for the forecasts of the conditional expectation of the students' weights. The length of interval for the actual forecast is greater than that of the conditional expectation given the same value of x_{n+1}, while the difference is not $2S_e$.

If x_{n+1} is equal to the previous sample mean $\bar{x}$, and if n is large, then the standard deviation of actual value, $\hat{y}_{n+1,i}$, as indicated in Eq. 14.19, approaches the standard error of the estimate s_e. This result should not be too surprising, for we know that the larger the sample, and the less a given value x_{n+1} deviates from $\bar{x}$, the more faith we have in the sampling results and in the subsequent forecast.

14.4.5 Using MINITAB to Calculate Confidence Interval and Interval

MINITAB output of Table 14.4 is presented in Fig. 14.9. In Fig. 14.9, $C_1 = x$, $C_2 = y$, fit $= \hat{y}_{n+1}$, standard deviation fit $= s(\hat{y}_{n+1})$, and MS of error $= s_e^2$. As discussed previously,

$$s(\hat{y}_{n+1,i}) = \sqrt{s_e^2 + s^2(\hat{y}_{n+1})}$$

Finally, for $x = 61$, the fit, standard deviation fit, 95 % confidence interval (C.I.), and 95 % prediction interval (P.I.) are presented in the last row of Fig. 14.9.

Example 14.1. Forecasting the Average Number of Cars in a Household of Three People We return to the data on cars and people per household similar to Example 13.4 of Chap. 13 to forecast the average number of cars in a household of 3 people. These data are given in Table 14.5.

From Table 14.5, we can obtain the following statistics:

$$\bar{y} = 2.4 \qquad \bar{x} = 3.8 \qquad \Sigma xy = 103 \qquad \Sigma x^2 = 162 \qquad \Sigma y^2 = 68a$$

$$\Sigma(x - \bar{x})(y - \bar{y}) = 11.8 \qquad \Sigma(x - \bar{x})^2 = 17.6 \qquad \Sigma(y - \bar{y})^2 = 10.4$$

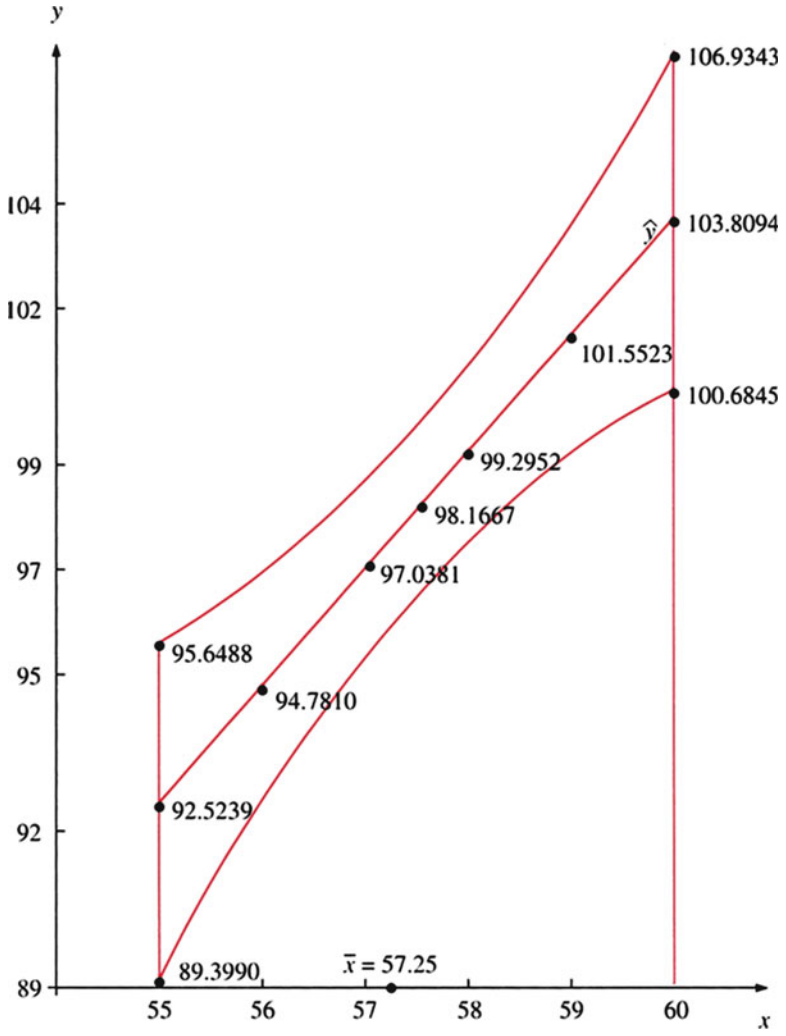


Fig. 14.7 Confidence interval for mean response with $x_{n+1} = \bar{x}$

We estimate the intercept and slope as

$$b = \frac{11.8}{17.6} = .67$$
$$a = 2.4 - (.67)(3.8) = -.146$$
$$\hat{y} = -.146 + .67x$$

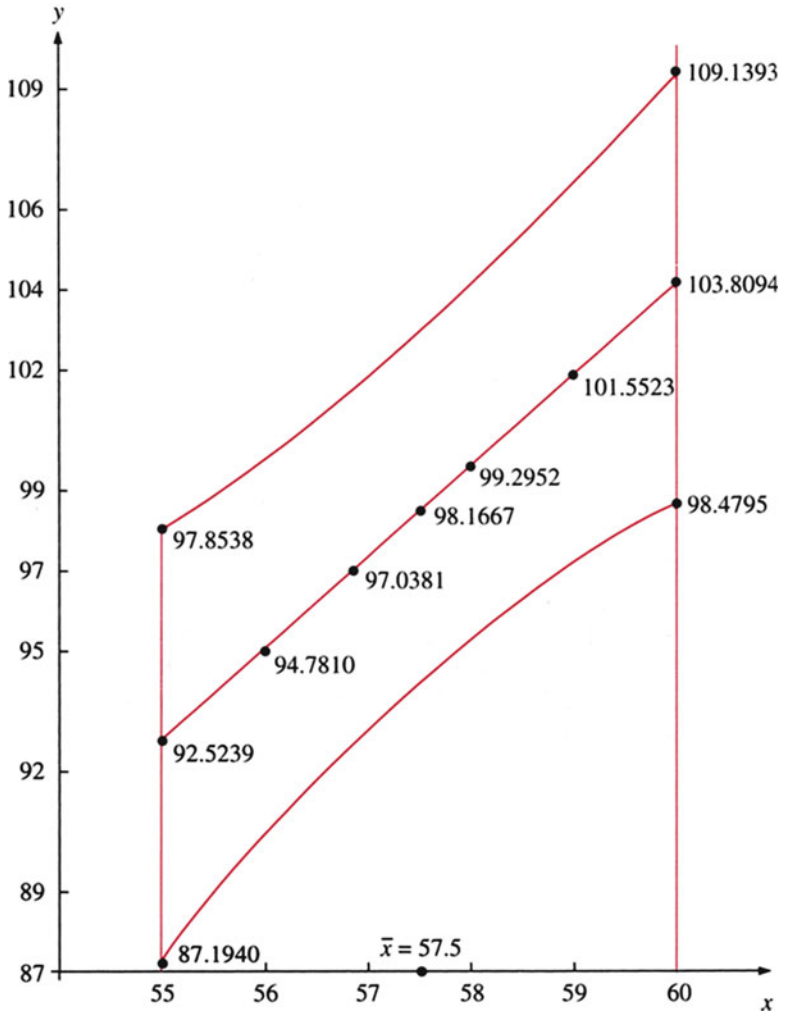


Fig. 14.8 Prediction interval for individual response

The standard error of slope can be calculated as follows:

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} = \sqrt{\frac{2.489}{8}} = .5578$$

$$\text{Standard error of } b = \frac{s_e}{\sqrt{\sum (x - \bar{x})^2}} = \frac{.5578}{\sqrt{17.6}} = .1329$$

```
MTB > READ C1 C2
DATA> 55 92
DATA> 56 95
DATA> 57 99
DATA> 58 97
DATA> 59 102
DATA> 60 104
DATA> END
      6 rows read.
MTB > REGRESS C2 1 C1;
SUBC> DW;
SUBC> PREDICT 61.
```

Regression Analysis

```
The regression equation is
C2 = - 31.6 + 2.26 C1

Predictor       Coef        StDev           T         P
Constant      -31.62        21.39       -1.48     0.213
C1            2.2571       0.3718        6.07     0.004

S = 1.555       R-Sq = 90.2%     R-Sq(adj) = 87.8%

Analysis of Variance

Source        DF           SS           MS          F         P
Regression     1       89.157       89.157      36.86     0.004
Error          4        9.676        2.419
Total          5       98.833

Durbin-Watson statistic = 3.03

     Fit  StDev Fit        95.0% CI              95.0% PI
 106.067      1.448   ( 102.045, 110.088)   ( 100.165, 111.968)
```

Fig. 14.9 MINITAB output of Table 14.4

Dividing .67 by .1329, we obtain the t statistic for b.

$$t = \frac{.67}{.1329} = 5.041$$

From Table A4 in Appendix A, we obtain $t_{8,.025} = 2.306$. Because 5.041 is larger than 2.306, we conclude that the estimated b is significantly different from zero at the 95 % level of significance. A family of 3 people will have an average of 1.864 cars [$-.146 + .67(3) = 1.864$]. The 95 % confidence interval for the average

Table 14.5 Numbers of cars per household

Household	Cars	People
1	4	6
2	1	2
3	3	4
4	2	3
5	2	4
6	3	4
7	4	6
8	1	3
9	2	2
10	2	4
Total	24	38

number of cars in a family of 3 people is constructed in accordance with Eq. 14.21 as

$$\hat{y}_{n+1} \pm t_{n-1,\alpha/2} s(\hat{y}_{n+1})$$

where

$\hat{y}_{n+1}$ = the mean value of y at the $(n + 1)$th level of x

$$\begin{aligned} s(y_{n+1}) &= s_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x - \bar{x})^2}} \\ &= (.5578)\sqrt{\frac{1}{10} + \frac{(3 - 3.8)^2}{17.6}} \\ &= (.5578)(.3693) = .2060 \end{aligned}$$

The 95 % confidence interval is

$$1.864 \pm (2.306)(.2060) = 1.864 \pm .4750 = 1.389 \text{ to } 2.339$$

On the basis of sample data, we are 95 % confident that the average number of cars in a household of 3 people will be in the interval of 1.389 cars to 2.339 cars.

14.5 Business Applications

In this section, we employ data on stock rates of return, auditing data, and other business data to show how simple linear regression and correlation analyses can be used in various real-world business applications.

Table 14.6 Subsidy rate and layoff rate for 11 industries

Industry	Subsidy rate (%), x	Layoff rate, y
Apparel	57	12.54
Chemicals	32	1.78
Construction	31	7.10
Electrical machinery	29	8.38
Fabricated metals	27	11.72
Food	36	5.10
Machinery	32	4.44
Misc. manufacturing	61	9.82
Primary metals	23	7.34
Retail	27	1.98
Wholesale trade	33	1.86

Source: Tropel, R.H.: On layoffs and unemployment insurance. Am. Econ. Rev. **83**, 541–559 (1983). Reprinted by permission of the publisher

Application 14.1 The Relationship Between Layoff Rate and the Unemployment Compensation Subsidy Rate To test whether the unemployment compensation subsidy causes firms to lay off more people than they would if they knew that they would not receive an outside subsidy for layoffs, Tropel (1983) used the data of unemployment compensation subsidy rate x (as a percentage of total revenues) and the layoff rate y (number of workers per 1,000) to do regression analysis for 11 industries, as indicated in Table 14.6.

The model of regressing y against x can be denned as

$$y_i = a + bx_i + e_i$$

where y_i and x_i are subsidy rate and layoff rate for i^{it} industries, respectively. The MINITAB output of this regression is presented in Fig. 14.10. From Fig. 14.10, we find that the estimated slope $b = .14468$ and the t-value associated with this slope is 1.55. From Table A4 in Appendix A, we find $t_{9,.025} = 2.262$, which is larger than 1.55. Hence, we conclude that the subsidy rate does not contribute information for the prediction of the layoff rate at $\alpha = .05$.

Application 14.2 Market Model Estimation and Analysis In Chaps. 2 through 4, annual rates of return for Johnson & Johnson and annual market rates of return during 1990–2009 were analyzed in detail. Now let's see how the market rates of return are used to explain the variations in rates of return for JNJ. The regression relationship can be defined as

$$R_{g,t} = a + bR_{m,t} + e_{g,t} \tag{14.23}$$

where

$R_{g,t}$ = annual rate of return for JNJ in period t
$R_{m,t}$ = market rate of return in period t

```
MTB > READ C1 C2
DATA> 57 12.54
DATA> 32 1.78
DATA> 31 7.10
DATA> 29 8.38
DATA> 27 11.72
DATA> 36 5.10
DATA> 32 4.44
DATA> 61 9.82
DATA> 23 7.34
DATA> 27 1.98
DATA> 33 1.86
DATA> END
     11 rows read.
MTB > REGRESS C2 1 C1;
SUBC> DW.
```

Regression Analysis

```
The regression equation is
C2 = 1.45 + 0.145 C1

Predictor       Coef        StDev           T         P
Constant       1.448        3.471        0.42     0.686
C1           0.14468      0.09339        1.55     0.156

S = 3.625       R-Sq = 21.1%     R-Sq(adj) = 12.3%

Analysis of Variance

Source       DF          SS          MS          F         P
Regression    1       31.53       31.53       2.40     0.156
Error         9      118.24       13.14
Total        10      149.77

Durbin-Watson statistic = 1.74
```

Fig. 14.10 MINITAB regression output for Application 14.1

Equation 14.23 is called the *market model* in financial analysis. It is often used to investigate the relationship between rates of return for individual securities and market rates of return. Further implications of the market model will be discussed in Chap. 21.

First we use data listed in columns (2) and (3) of Table 14.7 to estimate the market model for JNJ. MINITAB output of the market model for JNJ is presented in Fig. 14.11. From Fig. 14.11, we can define the estimated sample regression line as

Table 14.7 Rates of return for JNJ and market rates of return (1990–2009)

Data display

Row	JNJ	S&P
1	0.230108	0.036396
2	0.616842	0.124301
3	−0.551293	0.105162
4	−0.091578	0.085799
5	0.244915	0.019960
6	0.584558	0.176578
7	−0.409758	0.237724
8	0.340804	0.302655
9	0.287688	0.242801
10	0.124207	0.222782
11	0.139719	0.075256
12	−0.431191	0.163282
13	−0.078010	0.167680
14	−0.021172	0.028885
15	0.248595	0.171379
16	−0.032496	0.067731
17	0.122480	0.085510
18	0.034602	0.127230
19	−0.076435	0.174081
20	0.108473	0.222935

Descriptive statistics: JNJ

Total Variable	Count	Mean	SE mean	StDev	Variance	Minimum	Q1	Median
JNJ	20	0.0696	0.0677	0.3028	0.0917	0.5513	0.0776	0.1155

Variable	Q3	Maximum	Skewness	Kurtosis
JNJ	0.2477	0.6168	0.30	0.21

Descriptive statistics: S&P

Total Variable	Count	Mean	SE mean	StDev	Variance	Minimum	Q1	Median
S&P	20	0.0662	0.0338	0.1512	0.0229	0.2229	0.0167	0.0857

Variable	Q3	Maximum	Skewness	Kurtosis
S&P	0.1753	0.3027	0.56	0.49

$$\hat{R}_{g,t} = .0273 + .639\, R_{m,t} \tag{14.23a}$$

The positive value for b implies that rates of return for JNJ increase when market rates of return increase. The $b = .639$ means that a 1 % increase in market rates of return, $R_{m,t}$, in 1 year is associated with an increase in the annual rate of return for JNJ during the next year of about .639 %.

Fig. 14.11 MINITAB output of market model for JNJ

Regression Analysis: JNJ versus S&P

The regression equation is
JNJ = 0.0273 + 0.639 S&P

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.02726	0.07227	0.38	0.710	
S&P	0.6386	0.4472	1.43	0.170	1.000

S = 0.294804 R-Sq = 10.2% R-Sq(adj) = 5.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.17720	0.17720	2.04	0.170
Residual Error	18	1.56437	0.08691		
Total	19	1.74158			

Obs	S&P	JNJ	Fit	SE Fit	Residual	St Resid
1	0.036	0.2301	0.0505	0.0673	0.1796	0.63
2	0.124	0.6168	0.1066	0.0709	0.5102	1.78
3	0.105	-0.5513	0.0944	0.0682	-0.6457	-2.25R
4	0.086	-0.0916	0.0821	0.0665	-0.1736	-0.60
5	0.020	0.2449	0.0400	0.0691	0.2049	0.71
6	0.177	0.5846	0.1400	0.0823	0.4445	1.57
7	0.238	-0.4098	0.1791	0.1011	-0.5888	-2.13R
8	0.303	0.3408	0.2205	0.1246	0.1203	0.45
9	0.243	0.2877	0.1823	0.1029	0.1054	0.38
10	0.223	0.1242	0.1695	0.0962	-0.0453	-0.16
11	0.075	0.1397	0.0753	0.0660	0.0644	0.22
12	-0.163	-0.4312	-0.0770	0.1220	-0.3542	-1.32
13	-0.168	-0.0780	-0.0798	0.1236	0.0018	0.01
14	-0.029	-0.0212	0.0088	0.0785	-0.0300	-0.11
15	0.171	0.2486	0.1367	0.0810	0.1119	0.39
16	0.068	-0.0325	0.0705	0.0659	-0.1030	-0.36
17	0.086	0.1225	0.0819	0.0665	0.0406	0.14
18	0.127	0.0346	0.1085	0.0713	-0.0739	-0.26
19	-0.174	-0.0764	-0.0839	0.1261	0.0075	0.03
20	-0.223	0.1085	-0.1151	0.1452	0.2236	0.87

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 2.51280
Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	0.0820	0.0665	(-0.0577, 0.2217)	(-0.5529, 0.7169)

To obtain the goodness-of-fit measures and other statistics, we need to calculate the components of total variation for $R_{g,t}$. The analysis of variance in Fig. 14.11 reveals that the total variation in $R_{g,t}$ is

$$\text{SST} = \sum_{t=1}^{20} (R_{g,t} - \bar{R}_g)^2 = 1.74158$$

This SST can be decomposed as

$$\begin{aligned}\text{Explained variation (SSR)} &= \sum_{t=1}^{20} (\hat{R}_{g,t} - \bar{R}_{g,t})^2 \\ &= .17720\end{aligned}$$

and

$$\begin{aligned}\text{Unexplained variation (SSE)} &= \sum_{t=1}^{20} (R_{g,t} - \hat{R}_{g,t})^2 = \sum_{t=1}^{20} e_{g,t}^2 \\ &= 1.56437\end{aligned}$$

Drawing on our information about SST, SSR, SSE, and the related degrees of freedom, we construct the ANOVA table presented in Table 14.8.

Substituting information indicated in Tables 14.7 and 14.8 into Eqs. 13.20, 13.21, 14.2, 14.4, and 14.9, we obtain

$$\begin{aligned}s_e &= \sqrt{\sum_{t=1}^{n} e_{g,t}^2 / n - 2} \\ &= \sqrt{1.56437/18} = .2948 \\ R^2 &= \frac{.17720}{1.74158} = .102 \\ s_b &= s_e \Big/ \sqrt{\sum_{t=1}^{20} (R_{m,t} - \bar{R}_m)^2} \\ &= .2948/.6592 = .4472 \\ t_{18} &= .6386/.4472 = 1.43 \\ F_{1,18} &= \frac{\text{SSR}}{\text{SSE}/18} = \frac{.17720}{.08691} \\ &= 2.04\end{aligned}$$

Table 14.8 ANOVA for Eq. 14.23

Sources of variation	Sum of squares	Degrees of freedom	Mean square
Regression	0.17720	1	0.17720
Error	1.56437	18	0.08691
Total	1.74158	19	0.09166

The standard error of residuals, $s_e = .2948$, can be used to measure the absolute goodness of fit for the estimated market model, $\hat{R}_{g,t}$. And the coefficient of determination $R^2 = .102$ can be used to measure the relative goodness of fit for $R_{g,t}$. The estimated R^2 implies that 10.2 % of the variation in rates of return for JNJ has been explained by the variation of market rates of return. $t_{18} = 1.43$ can be used to test the following null hypothesis:

$$\mathrm{H}_0 : \beta = 0 \text{ versus } \mathrm{H}_1 : \beta \neq 0$$

Using a significance level of $\alpha = 5$ %, we have $t_{.025,18} = 2.1010$. Because $2.1010 > 1.43$, we cannot reject the null hypothesis and conclude that there is a linear relationship between $r_{g,t}$ and $R_{m,t}$ at $\alpha = 5$ %.

Equation 14.23a can be used to forecast the rate of return for JNJ in 2010. If we assume that $R_{m,n+1} = .0857$ for 2010, then the forecasted rate of return for JNJ in 2010 is

$$\begin{aligned}\hat{R}_{g,2010} &= .0273 + (.639)(.0857)\\ &= .0820\end{aligned}$$

We use Eq. 14.19 to calculate the standard error prediction (s_p).

$$\begin{aligned}s_\rho^2 &= s_e^2\left(1 + \frac{1}{20} + \frac{\left(R_{m,n+1} - \bar{R}_m\right)^2}{\sum_{t=1}^{20}\left(R_{m,t} - \bar{R}_m\right)^2}\right)\\ &= (.0869)\left(1 + \frac{1}{20} + \frac{(.0857 - .0662)^2}{.4344}\right)\\ &= (.0869)(1.05087) = .09132\end{aligned}$$

where s_p represents the standard error of the forecasted rate of return for JNJ in 2010. Using this information, we can estimate a 95 % forecast interval for the rate

of return for JNJ in 2010. From Table A4 in Appendix A, the value of $t_{.025,18}$ is 2.1010. Using related information, we get the following 95 % interval estimate:

$$\begin{aligned} .0820 \pm (2.1010)\sqrt{(.09132)} &= .0820 \pm .6349 \\ &= (-.5529,\ .7169) \end{aligned}$$

The numerical interval (−.5529, .7169) is not expected to include the true value for 95% of the time. 95% only works for the situation prior to plug observed values into intervals. This prediction interval is almost identical to that obtained via MINITAB (Fig. 14.11).

Application 14.3 Relationship Between Audited and Book Inventory Value Accountants often use the audit sampling approach to do statistical auditing. Auditors use a simple regression model like that indicated in Eq. 14.24 to estimate the relationship between client-reported account values and audited account values:

$$y_i = a + bx_i + e_i \tag{14.24}$$

where

y_i = ith audited account value
x_i = ith reported account value

Using as an example the inventory valuation demonstration data indicated in Table 14.9, we will show how regression analysis can be used to estimate the mean per-unit audited account value of the client population (μ_y) as defined in Eq. 14.25.

$$\hat{\mu}_y = a + b\mu_x = \bar{y} + b(\mu_x - \bar{x}) = \bar{y} - b\bar{x} + b\mu_x \tag{14.25}$$

where

$\bar{y}$ = mean of y_i
$\bar{x}$ = mean of x_i

μ_x = mean per-unit reported account value (a population parameter)

The use of Eq. 14.25 is similar to the case of mean response with $x_{n+1} \neq \bar{x}$ in the last section, constructing interval estimates of forecasts.

The 30 sample inventory accounts are randomly drawn from a population with the following population information:

Number of accounts = N = 2,000
Reported aggregate account value = $\sum X$ = $400,000
Mean per-unit reported account value = μ_x = $200

From the data listed in Table 14.9, we obtain

$$b = .9122 \text{ and } s_e = \$8.8180$$

Table 14.9 Inventory valuation demonstration data

(1)	(2)	(3)	(4)
Sample item number, n_i	Account number	Reported account value, x_i	Audited account value, y_i
1	2545	$ 161.21	$ 168.69
2	3988	183.68	174.53
3	3825	246.80	255.70
4	2613	207.28	208.46
5	3071	169.52	180.12
6	2848	180.26	189.76
7	3207	221.28	227.55
8	2109	185.58	174.61
9	2299	236.34	243.62
10	3052	202.44	209.35
11	2486	184.76	198.66
12	2822	191.21	198.51
13	3818	198.86	219.76
14	3674	192.65	208.46
15	2304	210.83	214.12
16	3206	208.59	219.41
17	3659	205.98	215.83
18	3544	148.35	172.39
19	3666	197.77	192.84
20	3937	238.25	249.08
21	3187	244.85	231.89
22	2622	192.28	191.72
23	2530	179.93	172.80
24	2320	180.81	187.11
25	2943	194.53	192.32
26	3670	216.40	221.92
27	3506	201.34	219.25
28	2416	212.00	204.39
29	2135	190.21	201.51
30	3181	188.29	205.40
		$5,972.28	$6,149.76
		$\bar{x} = \$199.076$	$\bar{y} = \$204.992$

Source: Bailey A.D. Jr.: Statistical Auditing: Review, Concepts and Problems, pp. 124–125. San Diego, Harcourt (1981). Copyright © 1981 by Harcourt Brace Jovanovich, Inc., reprinted by permission of the publisher

Substituting $\hat{b} = .9122$, $\bar{y} = \$204.992$, $\bar{x} = \$199.076$, and $\mu_x = \$200.00$ into Eq. 14.25, we have

$$\begin{aligned}\hat{\mu}_y &= \$204.992 + (.9122)(\$200.00 - \$199.076) \\ &= \$205.8349\end{aligned}$$

Following Eq. 14.21, we can construct the confidence interval associated with μ_Y as

$$\hat{\mu}_y - t_{(\alpha/2,n-2)} s_e \left[\frac{1}{n} + \frac{(\mu_x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^{1/2} < \mu_Y$$
$$< \hat{\mu}_y + t_{(\alpha/2,-2)} S_e \left[\frac{1}{n} + \frac{(\mu_x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^{1/2} \tag{14.26}$$

Let $\alpha = .10$, then $t_{.05,28} = 1.701$. Substituting $\bar{Y}, s_e, n, \mu_x, \bar{x}, \Sigma(x_i - \bar{x})^2$ and $t_{.05,28}$ into Eq. 14.26, we have

$$\$205.8349 - (1.701)(1.6113) < \mu_Y < \$205.8349 + (1.701)(1.6113)$$
$$\$203.0941 < \mu_Y < \$208.5757$$

We are 90% confident that the interval determined will include the true value of the per-unit audited inventory account value.

Application 14.4 Hamburger Sales: Predicting Profits. Healthy Hamburgers has a chain of 12 stores in Northern Illinois.[5] Sales figures and profits for the stores are given in Table 14.10. Our task is to obtain a regression line for the data and, assuming sales of $10 million, to predict profit for one store.

A worksheet for calculating regression coefficients is presented in Table 14.11. Substituting data from Table 14.11 into Eqs. 13.12 and 13.11, we obtain slope and intercept estimates.

$$b = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} = \frac{12(35.29) - 132(2.7)}{12(1796) - (132)^2} = .01593$$
$$a = \frac{\sum y_i - b(\sum x_i)}{n} = \frac{2.71 - .01593(132)}{12} = .0506$$

Thus, the regression line can be defined as

$$\hat{y} = .0506 + .01593x$$

[5] This application is drawn from Stevenson W.J.: Production/Operations Management, 2nd ed., pp. 137–141. Homewood, Irwin (1986). Reprinted by permission of Richard D. Irwin.

Table 14.10 Sales and profit for 12 Healthy Hamburgers stores

Sales, x_i (millions)	Profits, y_i (millions)
$7	$0.15
2	0.10
6	0.13
4	0.15
14	0.25
15	0.27
16	0.24
12	0.20
14	0.27
20	0.44
15	0.34
7	0.17

For sales of $x = 10$ (i.e., $10 million), estimated profit is

$$\hat{y}_{n+1} = .0506 + .01593(10) = .2099, \text{ or } \$209,900$$

Here we estimate standard error of the estimate in accordance with Eq. 13.35.

$$\sqrt{\frac{\sum y_i^2 - (\sum y_i)^2/n - b^2\left[\sum x_i^2 - (\sum x_i)^2/n\right]}{n-2}}$$

$$s_e = \sqrt{\frac{.7159 - (2.71)^2/12 - (.01593)^2\left[(1796) - \frac{(132)^2}{12}\right]}{12-2}}$$

$$= .04074, \text{ or } \$40{,}740$$

The prediction interval of $\hat{y}_{n+1,i}$ for the given value of x (i.e., $x_{n+1,i}$) can be obtained from Eq. 14.22.

If we substitute $s_e = .04074$, $n = 12$, $x_{n+1} = 10$, $\Sigma x^2 = 1796$, and $\Sigma x = 132$ into the standard error of Prediction (s_{pred}) portion of Eq. 14.22, we obtain

$$s_{\text{pred}} = (.04074)\sqrt{1 + \frac{1}{12} + \frac{(10-11)^2}{1796 - (132)^2/12}} = .04245$$

Table 14.11 Worksheet for calculating regression coefficients for Healthy Hamburgers

x	y	xy	x^2	y^2
1	0.15	1.05	49	0.0225
2	0.10	0.20	4	0100
6	0.13	0.78	36	0.0169
4	0.15	0.60	16	0.0225
14	0.25	3.50	196	0.0625
15	0.27	4.05	225	0.0729
16	0.24	3.84	256	0.0576
12	0.20	2.40	144	0.0400
14	0.27	3.78	196	0.0729
20	0.44	8.80	400	0.1936
15	0.34	5.10	225	0.1156
7	0.17	1.19	49	0.0289
132	2.71	35.29	1796	0.7159

Because $t_{12-2,.05/2} = 2.23$ (from Table A4 in Appendix A), a 95 % confidence interval for predicted y, y_{n+1}, is

$$\begin{aligned}\hat{y}_{n+1,i} \pm t_{n-2,\alpha/2}(s_{\text{pred}}) &= .2099 \pm 2.23(.04245) \\ &= .2099 \pm .0947\end{aligned}$$

or

$$.1152 \text{ to } .3046$$

That is, estimated profit on sales of \$10 million is \$209,900, and on the basis of sample data, we are 95 % confident that actual profit will be in the range of \$115,200–\$304,600.

Now we will test whether the estimated slope $b = .01593$ is significantly different from zero. We can estimate the standard deviation of b in accordance with Eq. 14.2 as

$$\begin{aligned}s_b &= s_e / \sqrt{\sum x^2 - \left(\sum x\right)^2 / n} \\ &= (.04074) \sqrt{\frac{1}{1796 - (132)^2/12}} = .0022\end{aligned}$$

By dividing .0022 into .01593, we obtain the t-value for b.

Table 14.12 Percentage of US TV households (x) and network share of TV revenues (y) during 1980–1985

	US cable TV households (%), x	Network share of television revenues, y
1980	21.1	98.9
1981	23.7	97.9
1982	25.8	96.5
1983	30.0	95.2
1984	35.7	94.0
1985	41.1	91.9

Source: Krugman D.M., Rust R.T.: The impact of cable penetration on network viewing. J. Mark. Res. **27**(9), 9–12 Oct./Nov. 1987

$$t_{12-2} = \frac{b}{s_b} = \frac{.01593}{.0022} = 7.24$$

Table A4 in Appendix A reveals that $t_{10,.025} = 2.23$. Because 7.24 is larger than 2.23, we can conclude that there is a strong relationship between the two variables (profit and sales) for 12 Healthy Hamburgers stores.

Application 14.5 The Impact of Cable TV Penetration on Network Share TV Revenues To investigate the effect of cable TV penetration on network share of advertising revenue, Krugman and Rust (1987) used the data listed in Table 14.12 to do the following regression analysis:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where

y_t = network share of TV advertising revenue in period t
x_t = the percentage of US TV households subscribing to cable TV in period t

The MINITAB regression output is presented in Fig. 14.12. From this output the estimated simple linear regression can be written as

$$\hat{y}_i = 106 - .335x_t$$
$$(147.22)\ (-14.18)$$

t-values are presented in the parentheses.

From Table A4 in Appendix A, we find that $t_{4,.005} = 4.604$. This critical value of t is smaller than both 147.22 and 14.18; hence, we can conclude that both the estimated intercept and the slope are significantly different from 0 at $\alpha = .01$.

Finally, we find that the confidence interval for prediction is (85.885, 89.218) and the prediction interval for $x = 54$ is (85.544, 89.559).

```
MTB > READ C1 C2
DATA> 21.1 98.9
DATA> 23.7 97.9
DATA> 25.8 96.5
DATA> 30.0 95.2
DATA> 35.7 94.0
DATA> 41.1 91.9
DATA> END
      6 rows read.
MTB > BRIEF 1
MTB > REGRESS C2 1 C1;
SUBC> DW;
SUBC> PREDICT 54.

Regression Analysis

The regression equation is
C2 = 106 - 0.335 C1

Predictor       Coef        StDev           T         P
Constant     105.634        0.718      147.22     0.000
C1          -0.33486      0.02362      -14.18     0.000

S = 0.4030      R-Sq = 98.0%     R-Sq(adj) = 97.6%

Analysis of Variance

Source        DF           SS           MS          F         P
Regression     1       32.644       32.644     200.98     0.000
Error          4        0.650        0.162
Total          5       33.293


     Durbin-Watson statistic = 1.69

          Fit  StDev Fit        95.0% CI              95.0% PI
       87.551      0.600    (  85.885,  89.218)  (  85.544,  89.559) XX
     X  denotes a row with X values away from the center
     XX denotes a row with very extreme X values
```

Fig. 14.12 MINITAB output for Application 14.5

14.6 Using Computer Programs to Do Simple Regression Analysis

In general, regression analysis is done on an electronic computer. Without the help of a computer, the arithmetic involved would be very time-consuming.

Most modern computing facilities have available prewritten computer program packages such as the MINITAB, the SAS, the SPSS, and the BMDP for carrying out regression analysis. These packages are available for use with personal computers. The user need only input the data and specify the model that is to be fitted for empirical analyses. All these packages can produce all the information we discussed in this chapter and in Chap. 13. In addition, these packages enable the

user to select options that produce much more numerical and graphical output. In Chaps. 13 and 14, we have applied the MINITAB to run simple linear regression. Now we show how the SAS program can be used to run simple linear regression.

Drawing on a set of sample market sales data called "Territory data for Click ballpoint pens" (see Table 14.13), we will show how the SAS computer program can be used to do the simple regression analysis discussed in this and the last chapter. This set of data represents the annual territory sales of Click, a national manufacturer of ball point pens, and other related variables. The company intends to use this set of data to investigate the effectiveness of the firm's marketing efforts. The company uses regional wholesalers to distribute Click pens, and it supplements their efforts with company sales representatives and spot TV advertising. The data to be analyzed are sales (y), advertising (x_1), number of sales representatives (x_2), and wholesaler efficiency index (x_3), where 4 = outstanding, 3 = good, 2 = average, and 1 = poor.

First we input all data listed in Table 14.11 into the SAS regression program. Then we specify the models to be analyzed.

$$y_i = a_0 + a_1 x_{1i} + e_i, \quad i = 1, 2, \ldots, 40 \tag{14.27}$$

$$y_i = b_0 + b_1 x_{2i} + e_i, \quad i = 1, 2, \ldots, 40 \tag{14.28}$$

$$y_i = c_0 + c_1 x_{3i} + e_i, \quad i = 1, 2, \ldots, 40 \tag{14.29}$$

After we specify these three simple regression models on the SAS regression programs, we can have the SAS program do three simple regression analyses. Their outputs are presented in Figs. 14.13 and 14.14. Figure 14.13 presents the scatter diagrams; Fig. 14.14a, b present the regression outputs.

Figure 14.13a–c are scatter diagrams showing how (a) y and x_1 (b) y and x_2, and (c) y and x_3 are related. Figures 14.4a and 14.14b show the estimated regression coefficients a_1, b_1, and c_1 with t statistics 11.43, 11.524 and .012, respectively. From Table A3 in Appendix A, using interpolation, we find that $t_{38,.025} = 2.025$; we can conclude that both a_1 and b_1 are significantly different from zero at $\alpha = .05$. However, c_1 is not significantly different from zero at $\alpha = .05$.

By using the output listed in Figs 14.13 and 14.14, we can do related analyses in terms of the concepts and methodologies we learned in Chaps. 13 and 14.

14.7 Summary

In this chapter and Chap. 13, we used simple regression analysis and correlation analysis to determine the relationship between two variables. In Chap. 13, we discussed estimation of the intercept and slope. To determine whether the regression does a good job of explaining the dependent variable, we investigated in detail

Table 14.13 Territory data for Click ballpoint pens

Territory	Sales, Y (thousands)	Advertising, X_1 (TV spots per month)	Number of sales representatives, X_2	Wholesaler efficiency index, X_3
005	260.3	5	3	4
019	286.1	7	5	2
033	279.4	6	3	3
039	410.8	9	4	4
061	438.2	12	6	1
082	315.3	8	3	4
091	565.1	11	7	3
101	570.0	16	8	2
115	426.1	13	4	3
118	315.0	7	3	4
133	403.6	10	6	1
149	220.5	4	4	1
162	343.6	9	4	3
164	644.6	17	8	4
178	520.4	19	7	2
187	329.5	9	3	2
189	426.0	11	6	4
205	343.2	8	3	3
222	450.4	13	5	4
237	421.8	14	5	2
242	245.6	7	4	4
251	503.3	16	6	3
260	375.7	9	5	3
266	265.5	5	3	3
279	620.6	18	6	4
298	450.5	18	5	3
306	270.1	5	3	2
332	368.0	7	6	2
347	556.1	12	7	1
358	570.0	13	6	4
362	318.5	8	4	3
370	260.2	6	3	2
391	667.0	16	8	2
408	618.3	19	8	2
412	525.3	17	7	4
430	332.2	10	4	3
442	393.2	12	5	3
467	283.5	8	3	3
471	376.2	10	5	4
488	481.8	12	5	2

Source: C. A. Gilbert, in G. A. Churchill, Jr., Marketing *Research: Methodological Foundations,* 3rd ed., 1983, p. 563. Copyright © 1983 by the Dryden Press, reprinted by permission of the publisher

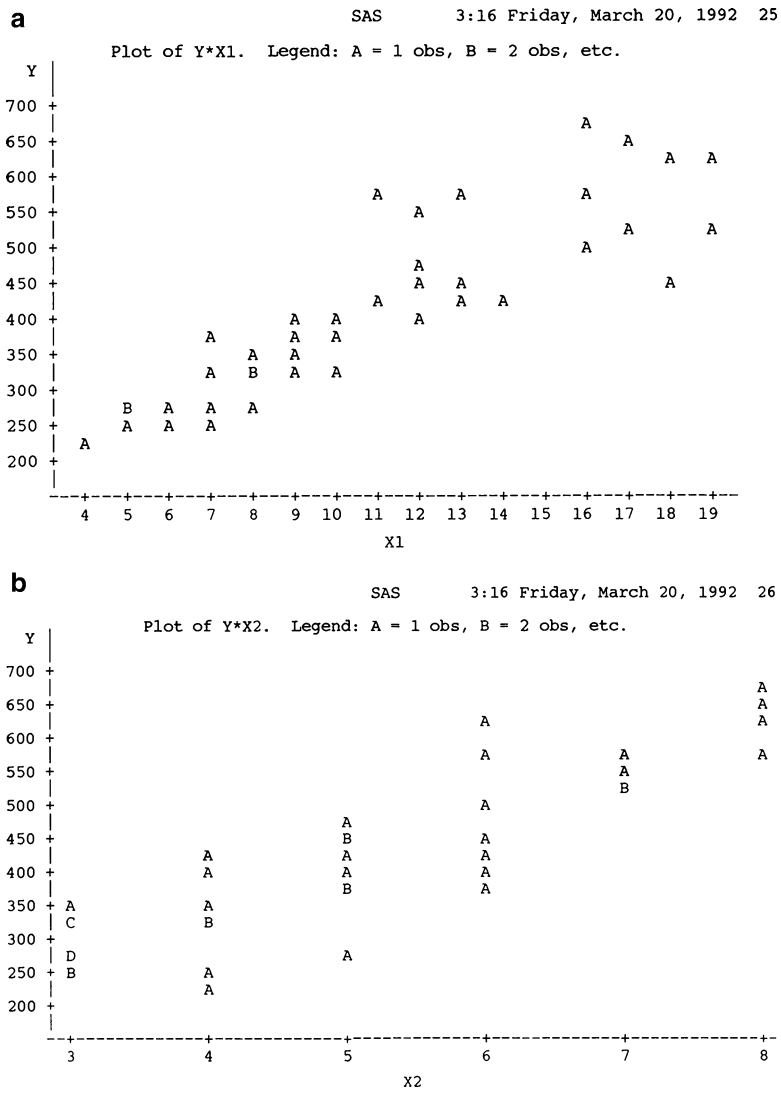


Fig. 14.13 (continued)

two goodness-of-fit measures: the standard error of residuals and the coefficient of determination.

Another measure of association, the sample correlation coefficient, was discussed in Chap. 13, along with the relationship between the correlation coefficient r and the slope estimate b and that between r and the coefficient of determination R^2.

Chapter 14 showed how researchers can use the standard error and the parameter value to construct a test to determine whether the parameter values are equal to 0.

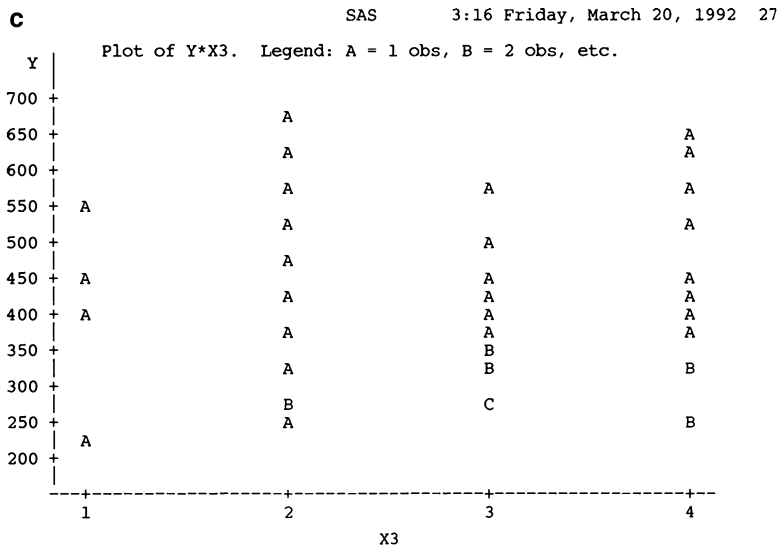


Fig. 14.13 (**a**) Scatter diagram showing the relationship between y and x_1 (**b**) Scatter diagram showing the relationship between y and x_2 (**c**) Scatter diagram showing the relationship between y and x_3

We also discussed the F-test and confidence intervals and point estimates for forecasting. The MINITAB program is used to do this kind of analysis. Applications of regression analysis drawn from finance, accounting, and marketing rounded out the picture. Finally, we saw how SAS statistical computer programs can be used to do simple regression analysis.

Questions and Problems

1. Here x is the number of units of a product produced during a certain period, and y represents total variable costs incurred during the period.

x	0	1	2	3	4	5	6
y	1	2	3	5	8	11	12

   (a) Find the estimated equation for the regression of y on x.
   (b) Find the predicted value of y given $x = 8$.
   (c) Find the standard error of estimate.
   (d) Find the coefficient of determination r^2.

2. Use again the data given for question 1.

   (a) Find the standard deviation of the regression line's slope s_b.
   (b) Find an interval that you can be 95 % confident will contain b, the slope of the population regression line.

a

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	465161.12840	465161.12840	132.811	0.0001
Error	38	133091.89535	3502.41830		
C Total	39	598253.02375			

Root MSE	59.18123	R-square	0.7775
Dep Mean	411.28750	Adj R-sq	0.7717
C.V.	14.38926		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	80.065802	30.22585829	2.649	0.0117
X2	1	66.244340	5.74818946	11.524	0.0001

Durbin-Watson D	2.041
(For Number of Obs.)	40
1st Order Autocorrelation	-0.040

Fig. 14.14 (continued)

3. In a regression problem, $n = 30$, $\Sigma x_i = 15$, $\Sigma y_i = 30$, $\Sigma x_i y_i = 30$, $\Sigma x_i^2 = 10$, and $\Sigma y_i^2 = 160$.

   (a) Find the regression line $\hat{y} = a + bx$.
   (b) Estimate the variance σ_{yx}^2.
   (c) Test H_0: $b = 0$ against H,: $b \neq 0$. Let $\alpha = .05$.

4. In a regression analysis, $n = 25$, $\Sigma X_i = 75$, $\Sigma Y_i = 50$, $\Sigma X_i^2 = 625$, $\Sigma X_i Y_i = 30$, and $\Sigma Y_i^2 = 228$.

   (a) Find the regression equation.
   (b) Find s_{yx}^2, s_a^2, and s_b^2.
   (c) Test whether $b = 0$. Let $\alpha = .01$. Use H_1: $b \neq 0$.
   (d) Find a 95 % confidence interval for a.

5. For each of the following sets of quantities, find the sample correlation coefficient. Test the hypothesis that $\rho = 0$. Let $\alpha = .05$. Use H_1: $\rho \neq 0$.

   (a) $n = 11$, $\Sigma y^2 = 400$, $\Sigma(x - \bar{x})(y - \bar{y}) = 400$, $\Sigma x^2 = 625$
   (b) $n = 18$, $\Sigma y^2 = 100$, $\Sigma(x - \bar{x})(y - \bar{y}) = 36$, $\Sigma x^2 = 36$

6. Find the sample correlation coefficient and test H_0: $\rho = 0$ for the data of question 4. Use H_1: $\rho \neq 0$ and $\alpha = .05$.

b

```
Dependent Variable: Y
                          Analysis of Variance

                            Sum of          Mean
      Source          DF      Squares       Square      F Value       Prob>F

      Model            1 465161.12840 465161.12840      132.811       0.0001
      Error           38 133091.89535   3502.41830
      C Total         39 598253.02375

          Root MSE      59.18123     R-square       0.7775
          Dep Mean     411.28750     Adj R-sq       0.7717
          C.V.          14.38926


                       Parameter Estimates

                    Parameter       Standard    T for H0:
     Variable  DF      Estimate          Error   Parameter=0    Prob > |T|

     INTERCEP   1     80.065802   30.22585829         2.649        0.0117
     X2         1     66.244340    5.74818946        11.524        0.0001


Durbin-Watson D             2.041
(For Number of Obs.)           40
1st Order Autocorrelation  -0.040


Dependent Variable: Y
                          Analysis of Variance

                            Sum of          Mean
      Source          DF      Squares       Square      F Value       Prob>F

      Model            1      2.19822      2.19822        0.000       0.9906
      Error           38 598250.82553  15743.44278
      C Total         39 598253.02375

          Root MSE     125.47288     R-square       0.0000
          Dep Mean     411.28750     Adj R-sq      -0.0263
          C.V.          30.50734


                       Parameter Estimates

                    Parameter       Standard    T for H0:
     Variable  DF      Estimate          Error   Parameter=0    Prob > |T|

     INTERCEP   1    410.606023   60.98903066         6.732        0.0001
     X3         1      0.241231   20.41491700         0.012        0.9906


Durbin-Watson D             1.724
(For Number of Obs.)           40
1st Order Autocorrelation   0.115
```

Fig. 14.14 (**a**) SAS output of $y_i = a + b_1x_{1i} + e_i$ (**b**) SAS Output of $y_i = a + b_2x_{2i} + e_i$ and $y_i = a + b_3x_{3i} + e_i$

7. Records were kept on the scores 14 job applicants got on a manual-dexterity test and on their production output after a week on the job.

   (a) Use the following data and MINITAB program to estimate the correlation between test score and production output.
   (b) Is there a significant correlation? Let $\alpha = .05$.

Score	112	72	61	50	48	117	13
Output	153	83	36	93	86	121	20
Score	19	13	43	84	31	124	66
Output	26	16	62	103	30	120	84

8. The managers of a weight-loss clinic wish to confirm their belief that there is a relationship between the weight of a person entering the program and the number of pounds lost. The table presents data for 8 of the clinic's clients.

x (beginning weight)	142	306	261	177
y (weight loss)	15	146	73	50
x (beginning weight)	205	165	289	154
y (weight loss)	25	15	36	12

Use the following regression information for the relationship between beginning weight and pounds lost to complete parts (a) and (b).

$$\hat{y} = -67.78 + .54x$$
$$r^2 = .58 \quad s_e = 31.60 \quad s_b = .186$$
$$\Sigma x^2 = 28,911.87 \quad \Sigma y^2 = 14,362.00$$
$$\Sigma xy = 15,557.48$$

(a) Test the hypothesis that there is no relationship between beginning weight and pounds lost, using the t-test on the slope of the regression line. Let $\alpha = .05$.

(b) Repeat the test in part (a), but use the t-test on r, the correlation coefficient.

9. A certain firm provides expense accounts for its executives. Using the following data, the firm's human resources department ran a regression analysis to determine whether a relationship exists between annual salary and amount claimed in expenses each year.

x (salary in \$ 1,000 s)	40	38	22	25	30	35	40
y (expenses in \$ 1,000 s)	0.8	1.6	0.6	0.9	1.2	2.2	1.1

The regression results were

$$\hat{y} = .227 + .030x \quad r = .398$$
$$s_{y|x} = .547 \quad s_b = .0306$$
$$\sum y^2 = 1.78 \quad \sum x^2 = 320.86 \quad \sum xy = 9.63$$

(a) Test the hypothesis that there is no relationship between annual salary and expenses claimed, using the t-test on the regression coefficient b. Use $\alpha = .05$.
(b) Repeat the test of part (a), but use the t-test on r.

10. Briefly explain what we mean when we say that α and β are unbiased. Why is unbiasedness an important property in an estimator?
11. Explain the purpose of constructing confidence intervals for the parameters α and β.
12. What assumptions do we make about the distributions of a and b? Why is this important in constructing confidence intervals?
13. When we are estimating the regression $y = \alpha + \beta x$, which parameter is of greater interest, α or β? Why?
14. In testing the significance of the parameters, why do we sometimes use a two-tailed test and sometimes use a one-tailed test?
15. In testing the significance of the parameters, why do we sometimes use a z-test and sometimes use a t-test?
16. Compare the width of a 90 % confidence interval with that of a 99 % confidence interval. Which is wider? Why?
17. Compare the standard error of the estimate discussed in Chap. 13 with the standard error of the regression coefficient discussed in this chapter.
18. What null hypothesis do we generally use in testing the significance of the slope coefficient b?
19. Ralph Farmer of the Department of Agriculture is interested in the relationship between the amount of fertilizer used and the number of bushels of wheat harvested. He collects the following information on six farmers.

x (pounds of fertilizer)	y (bushels of wheat)
100	1,000
150	1,250
180	1,710
200	2,100
222	2,500

Use MINITAB to do the following:

(a) Draw a scatter diagram for the data.
(b) Estimate the regression parameters for α and β.
(c) Calculate the standard error of the estimate and the standard error of the coefficient b.
(d) Calculate the t-value for the coefficient of b.
(e) If 210 lb of fertilizer is used, what amount of wheat can be expected to be harvested? What is the 95 % confidence interval? (Hint: Follow the procedure used in Fig. 14.10.)

20. Use the MINITAB output from question 19 to construct a 95 % confidence interval for b. Also construct a 90 % confidence interval for b. Which interval is larger?
21. Using the MINITAB output given in question 19, compute SST, SSE, SSR, and R^2. Use an F test to test the significance of the regression.
22. A recent study of Departments of Labor in all 50 states indicates that the amount (in thousands of dollars) spent on job placement for the unemployed and the number of people employed has a slope of 1.7 and a standard error of the regression coefficient of .43. Test the significance of the slope coefficient at the 95 % confidence level.

Use the following data and the MINITAB program to answer questions 23–30. The table gives monthly rates of return for 3-month treasury bills; the value-weighted New York Stock Exchange Index; and Chrysler, Ford, and GM stock.

	T-Bill	NYSE	Chrysler
Month	Rf	R_m	R_1
87.01	.004414	.12823	.29054
87.02	.004543	.04100	–.01309
87.03	.004543	.02469	.18037
87.04	.004583	–.01483	.03846
87.05	.004599	.00644	–.11111
87.06	004607	.04797	.01103
87.07	.004623	.04682	.19414
87.08	.004904	.03688	.09816
87.09	.005193	–.02085	–.06983
87.10	.004977	–.21643	–.35649
87.11	.004623	–.07547	–.23944
87.12	.004688	.06851	.10494

Month	Ford, R2	GM, R3
87.01	0.33378	0.14015
87.02	0.02689	0.00831
87.03	0.10475	0.04690
87.04	0.08741	0.15200
87.05	0.00137	–0.03889
87.06	0.08941	–0.03079
87.07	0.03.409	0.07564
87.08	0.06273	0.04923
87.09	0.09259	–0.09783
87.10	0.21939	–0.29518
87.11	0.05795	–0.01496
87.12	0.05975	0.08869

23. In finance, we are often interested in how the return of one stock is related to some market index such as the NYSE. The model we usually estimate to understand this relationship is known as the market model and is given by the equation

$$R_{j,t} = \alpha_j + \beta_j R_{m,t} + e_{j,t}$$

where

$R_{j,t}$ = return on stock j in month t
$R_{m,t}$ = return on some market index in month t
α_j = intercept of the regression line
β_j = slope of the regression line
$e_{j,t}$ = a random error term

Use MINITAB to do the following:

(a) Draw a scatter diagram for Ford and the NYSE index.
(b) Estimate the parameters α and β.
(c) Compute SSE, SSR, SST, R^2, and the standard error of the estimate for this regression.
(d) Compute the standard error for b, and use a t-test to test the significance of the slope of the regression.

24. In finance, we sometimes choose to estimate the capital asset pricing (CAPM) version of the market model, which is given by the equation

$$R_{j,t} - R_{f,t} = \alpha_j + \beta_j \left[R_{m,t} - R_{f,t}\right] + e_{j,t}$$

where $R_{j,t}$ is the return on a risk-free asset (such as T-bills) in month t. Repeat parts (a)–(d) of question 23 for the CAPM version of the market model using the MINITAB program. (The CAPM will be discussed in Chap. 21.)
25. Using R^2 as the measure of goodness of fit, compare the market model estimated in question 23 with the CAPM version estimated in question 24.
26. Find a 95 % confidence interval for the slope coefficients you calculated in questions 23 and 24. Which estimate of β has the wider confidence interval?
27. Suppose we are interested in testing whether β is equal to 1. Then we would test H_0: $\beta = 1$ against H_1: $\beta \neq 1$. Using the model given in question 23, test this hypothesis.
28. Repeat questions 23–27, using GM stock's rates of return.
29. Repeat questions 23–27, using Chrysler stock's rates of return.

30. Suppose we are interested in the relationship between the return on the risk-free asset (T-bills) and the return on the NYSE index.

    (a) Estimate the intercept and the slope for a regression of $R_{m,t}$ on $R_{f,t}$.
    (b) Compute the standard error of the regression and use a t-test to test the significance of b.
    (c) Calculate a 99 % confidence interval for b.

31. When we are interested in the relationship between a dependent variable and time, we sometimes use a time-trend regression. That is, we use a dependent variable that consists only of the day, month, or year of our observations. A time-trend regression is given by the equation

$$y_t = a + bt + e_t$$

where t represents time, $t = 1,2,3, \ldots, T$. Suppose we are interested in how Johnson & Johnson's inventory turnover has changed over time. We collect data on J&J's inventory turnover for a 20-year period from 1969 to 1988. Use the MINITAB program to answer the following:

    (a) Draw a scatter diagram for these data.
    (b) Estimate the regression coefficients α and β.
    (c) Calculate SSR, SSE, SST, and R^2, and the standard error of the estimate.
    (d) Compute the standard error of b, and use a t-test to test the significance of b.

t	J&J's inventory turnover	t	J&J's inventory turnover
1	3.19	11	2.71
2	3.02	12	2.70
3	2.96	13	2.78
4	3.10	14	2.38
5	2.92	15	2.28
6	2.28	16	2.37
7	2.77	17	2.45
8	2.76	18	2.33
9	2.84	19	2.27
10	2.76	20	2.32

32. Use the information and your calculations from question 31 to construct a 90 % and a 99 % confidence interval for b.

33. When estimating the relationship between the price of a good and the quantity of the good sold (the demand curve), economists sometimes choose to transform the price and quantity data by taking the natural logarithm of both. When this is done, the slope coefficient β can be interpreted as the price elasticity of demand (the sensitivity of quantity to changes in price). Consider the following information on the price and quantity of So-Good Candy Bars.

Price	Quantity
$1.50	100
1.25	135
1.00	175
0.75	225
0.50	300
0.25	500

Use the MINITAB program to answer the following:

(a) Estimate the elasticity of demand for these data.
(b) Use a t-test to test the significance of b.
(c) Construct a 95 % confidence interval for the price elasticity.

34. The batting instructor of the Minnesota Twins is interested in the relationship between number of hours of batting practice and batting average. He collects the following data on 8 players:

Hours of batting practice per week	Batting Average
5	0.265
8	0.277
9	0.254
10	0.320
11	0.301
9	0.260
7	0.230
6	0.272

Use the MINITAB program to answer the following:

(a) Draw a scatter diagram for these data.
(b) Compute the regression parameters α and β.
(c) Compute the standard error of b, and use a t-test to test the significance of the slope of the regression.
(d) Construct a 95 % confidence interval for b.

35. Suppose you estimate a regression of y against x and find that $b = 1.3$ and $\sigma_b = .4$ (population standard error of the regression, which is known).
    (a) Construct a 90 % and a 99 % confidence interval for b.
    (b) Now suppose σ_b is not known. How would this change the way you construct your confidence interval for b? Assume that 15 observations were used to estimate the regression.
36. Use the data from question 21 of Chap. 13 to compute the standard error of b. Use a t-test to test the significance of b. Construct a 99 % confidence interval for b.
37. Use the data from question 33 of Chap. 13 to compute the standard error of b. Use a t-test to test whether $b = 1$. Construct a 90 % confidence interval for b.
38. Use the data from question 27 of Chap. 13 to compute the standard errors of a and b. Construct a 95 % confidence interval for both a and b.
39. Use the information given in question 57 of Chap. 13, and construct a 95 % confidence interval for both a and b.
40. Investment advisors sometimes recommend holding gold as part of an investor's portfolio, because the value of gold appears to be negatively related to that of the stock market. Thus, when the stock market goes down in value, the value of gold goes up in value, and some of the investor's losses in the market are offset by gains in the value of her or his gold. The accompanying table shows data on annual rates of return for a gold mutual fund and for the S&P 500.

Year	Gold mutual fund	S&P 500
1979	151.30	18.16
1980	70.70	31.48
1981	−18.90	−4.85
1982	47.30	20.37
1983	8.20	22.30
1984	−25.30	5.97
1985	−11.00	31.05
1986	30.10	18.75
1987	51.50	5.24
1988	11.30	16.58

Use MINITAB to answer the following questions:

(a) Estimate the slope of the regression of the rates of return of the gold mutual fund against those of the S&P 500.
(b) Use a t-test to test the hypothesis that $b < 0$.
(c) If you expect the rate of return to be 20 % next year (1989), what is the rate of return of the gold mutual fund you expect next year? What is the 95 % confidence interval for your expectation?

41. Use the data from question 40 to construct a 95 % confidence interval for b. Is it possible for the true b to be negative?
42. Briefly explain how we can use regression analysis to forecast values of y.
43. Explain why we often construct interval estimates of forecasts.
44. How is the size of the forecast interval affected when we use values of x that are much greater or much less than the mean value of x for forecasting?
45. Use the data given in question 31 to forecast Johnson & Johnson's inventory turnover for 1989 and 1990. Construct a 95 % confidence interval for both of these forecasts. Use Eq. 14.21.
46. Use the regression estimated in question 40 to forecast the return for the gold mutual fund in 1989 and 1990. Assume that the best forecast for the return of the S&P 500 in 1989 and 1990 is the mean of the S&P 500's returns for the previous 5 years. Construct a 99 % confidence interval for both of these forecasts.
47. Use the data and the regression given in question 19 to forecast the number of bushels of wheat that will be harvested if 250 lb of fertilizer are used. Construct a 90 % and a 99 % confidence belt for the regression line. For which interval is the confidence belt wider? Explain.
48. Use the regression results from questions 23 and 24 to forecast the return for Ford in January 1988, using both the standard market model and the CAPM version of the market model. Assume that the return for the NYSE is 12 % in January 1988.
49. Construct a 95 % confidence interval for the forecasts produced in question 48. Which model has the larger interval? Use Eq. 14.21.
50. What is proxy error? Give some examples of proxy error in economics, accounting, and finance.
51. Briefly explain how proxy error of x affects the results from a standard linear regression.
52. Use your results from the regression given in question 65 of Chap. 13 to test the significance of b via a t-test. Also construct a 90 % confidence interval for b. (Hint: Calculations from question 66 in Chap. 13 also may help.)
53. Again using your results from question 65 of Chap. 13, forecast the value of J&J's current ratio. Assume that the best forecast of the industry's current ratio is the mean of that ratio. Construct a 99 % confidence interval for this forecast. (Hint: Your results from question 52 may be helpful.)
54. Redo questions 52 and 53, using J&J's inventory turnover.
55. Redo questions 52 and 53, using J&J's return on assets.
56. Redo questions 52 and 53, using J&J's price/earnings ratio.
57. Suppose you estimate the following simple regression:

$$\hat{y} = 50 + 2.23x$$
$$\text{SSE} = 22{,}300$$
$$n = 23$$
$$\sum (x - \bar{x})^2 = 2,700$$

(a) On the basis of the information provided, test the significance of the slope at the 99 % confidence level.
(b) Construct a 99 % confidence interval for the slope coefficient.

58. Suppose a researcher is interested in the relationship between the dollar volume of sales and the number of miles customers live from the store. She collects data on dollar volume of sales per customer (y) and the miles a customer lives from the store (x) for 28 customers. The following relationship is then estimated:

$$\hat{y} = 75 - .85x$$
$$s_b = .32$$
$$n = 28$$

(a) Interpret the meaning of the slope coefficient.
(b) Test whether the slope coefficient is significant at the 95 % level of confidence.

59. Using the information from question 58, construct a 90 % confidence interval for b.
60. In finance we are sometimes interested in hedging the risk associated with future price changes by using futures contracts. A futures contract allows the buyer of the contract to buy the commodity at a later date at a price that is agreed upon now. By purchasing the correct number of contracts, an investor can reduce or even eliminate his or her risk. The correct number of contracts to purchase is known as the hedge ratio, and it can be estimated by regression analysis. The regression to be estimated is

$$\Delta S = \alpha + \beta \Delta f + e$$

where

ΔS = change in the spot price of the commodity
Δf = change in the futures price of the commodity
β = hedge ratio (number of futures contracts used for hedging)

Suppose you collect 30 daily spot and future prices over a 1-year period and estimate β to be 3.32 with a standard error of 1.12. Construct a 95 % confidence

interval around the hedge ratio β. (See Appendix 2 of Chap. 19 for the derivation of hedge ratio.)

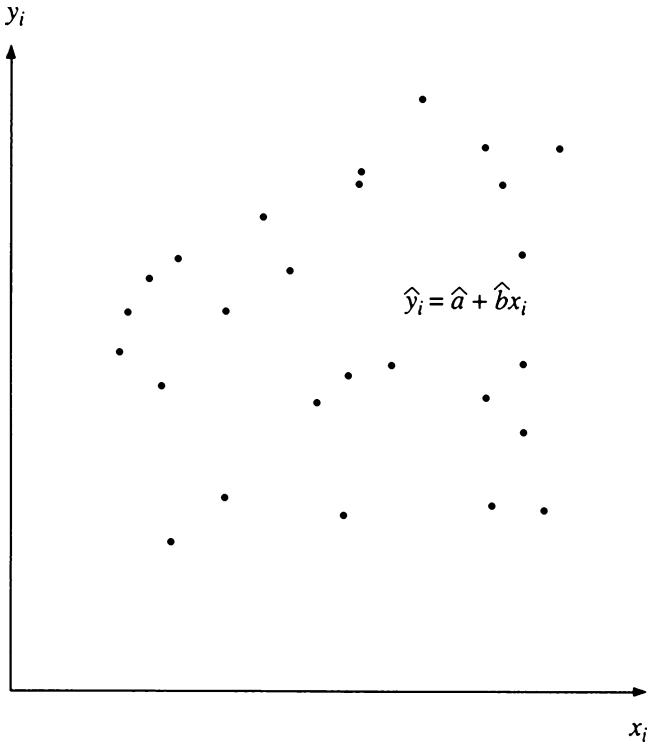
61. Estimate a simple regression model, using the least-squares method and the information given.

x	y	xy x^2 y^2 $\hat{y}$ e e^2 $(y - \bar{y})^2$
10	12.8	
20	18.9	
30	21	
40	38	
50	40	

62. Estimate a simple regression model, using the least-squares method and the information given.

x	y	xy x^2 y^2 y e e^2 $(y - \bar{y})^2$
10	52	
20	48	
30	31	
40	28	
50	10	

63. Evaluate the goodness of fit for the following graph:



64. The following table summarizes the sales X and advertising expenditures Y (both in millions) for Rivera Company.

X	10	20	30	40	50	60
Y	70	210	230	340	360	530

(a) Estimate the regression, using X as the explanatory variable.
(b) What will be the expected sales for next year if the company allocates \$70 million to advertising?
(c) Perform a test to see whether advertising expenditures have a positive impact on sales.

65. The table on page 649 lists the administrative and enrollment breakdowns for the schools of each municipality in Middlesex County, New Jersey. Using total enrollment as the independent variable and number of administrators as the dependent variable, run a simple regression by using information from both 1982–1983 and 1990–1991. Use the MINITAB or SAS programs.
66. The table below shows the undergraduate GPA and quantitative scores on the GRE of 10 students. Explain the MINITAB output on page 650.

Student	GPA	Quantitative scores on GRE
1	4.00	630
2	2.62	590
3	3.30	580
4	3.15	490
5	3.54	720
6	3.21	690
7	3.57	700
8	3.61	690
9	2.90	520
10	3.05	540

Administrative and enrollment breakdown

District	1982–1983 total enrollment	Number of administrators	1990–1991 total enrollment	Number of administrators
Carteret	2,962	29	2,525	28
Cranbury Twp	266	1	312	2
Dunellen Boro	901	5	826	6
East Brunswick Twp	7,652	36	6,657	43
Edison Twp	10,349	53	10,966	52
Highland Park Boro	1,625	10	1,441	13
Jamesburg Boro	497	3	410	2
Metuchen Boro	1,879	17	1,590	22
Middlesex Bor.	2,151	13	1,720	14
Middlesex Co-Ed Ser Comm	56	3	213	8
Middlesex City Vocational	4,181	19	3,314	27
Milltown Boro	708	4	628	5
Monroe Twp	2,545	19	2,485	22
New Brunswick City	4,286	33	4,086	32
North Brunswick Twp	3,319	25	3,996	24
Old Bridge Twp	9,120	48	8,037	50
Perth Amboy City	5,774	32	6,274	42
Piscataway Twp	6,155	38	5,637	39
Sayreville Boro	4,391	26	4,245	24
South Amboy City	903	8	948	7
South Brunswick Twp	3,125	20	3,871	37
South Plainfield Boro	3,381	24	3,001	20
South River Boro	1,650	10	1,502	10
Spotswood Boro	1,689	17	1,385	13
Woodbridge Twp	11,726	83	10,724	82
Middlesex County	91,291	576	86,793	624
Franklin	4,330	38	4,155	37

Source: Home News, December 15, 1991. Reprinted with permission of the publisher

MINITAB Output for Question 66

```
MTB > READ C1 C2
DATA> 4.00 630
DATA> 2.62 590
DATA> 3.30 580
DATA> 3.15 490
DATA> 3.54 720
DATA> 3.21 690
DATA> 3.57 700
DATA> 3.61 690
DATA> 2.90 520
DATA> 3.05 540
DATA> END
     10 rows read.
MTB > BRIEF 3
MTB > REGRESS C2 1 C1;
SUBC> DW;
SUBC> PREDICT C1.
```

Regression Analysis

```
The regression equation is
C2 = 229 + 117 C1

Predictor       Coef        StDev          T        P
Constant       229.2        201.4       1.14    0.288
C1            117.09        60.71       1.93    0.090

S = 72.65       R-Sq = 31.7%     R-Sq(adj) = 23.2%

Analysis of Variance

Source        DF          SS          MS         F        P
Regression     1       19630       19630      3.72    0.090
Error          8       42220        5278
Total          9       61850

Obs        C1         C2         Fit  StDev Fit   Residual    St Resid
  1      4.00      630.0       697.5       48.6      -67.5       -1.25
  2      2.62      590.0       536.0       47.0       54.0        0.98
  3      3.30      580.0       615.6       23.0      -35.6       -0.52
  4      3.15      490.0       598.0       24.6     -108.0       -1.58
  5      3.54      720.0       643.7       27.4       76.3        1.13
  6      3.21      690.0       605.0       23.5       85.0        1.24
  7      3.57      700.0       647.2       28.4       52.8        0.79
  8      3.61      690.0       651.9       29.9       38.1        0.58
  9      2.90      520.0       568.8       33.2      -48.8       -0.75
 10      3.05      540.0       586.3       27.4      -46.3       -0.69

Durbin-Watson statistic = 1.68

     Fit  StDev Fit        95.0% CI              95.0% PI
   697.5       48.6   (   585.5,   809.6)   (   496.0,   899.1)
   536.0       47.0   (   427.6,   644.3)   (   336.4,   735.5)
   615.6       23.0   (   562.6,   668.6)   (   439.8,   791.3)
   598.0       24.6   (   541.3,   654.8)   (   421.1,   774.9)
   643.7       27.4   (   580.6,   706.8)   (   464.6,   822.8)
   605.0       23.5   (   550.7,   659.4)   (   428.9,   781.2)
   647.2       28.4   (   581.7,   712.7)   (   467.3,   827.1)
   651.9       29.9   (   582.9,   720.8)   (   470.7,   833.1)
   568.8       33.2   (   492.1,   645.4)   (   384.5,   753.0)
   586.3       27.4   (   523.2,   649.4)   (   407.2,   765.4)
```

67. You are given the following information for x and y:

$\text{Cov}(x, y) = 6.3$	Mean of $y = 75$
$\text{Var}(x) = 4.2$	$\Sigma(x_i - \bar{x})^2 = 42$
$\text{Var}(y) = 2,727.7$	$s_e = 50$
Mean of $x = 100$	$n = 12$

(a) Compute the least-squares estimates for the slope and intercept.
(b) Compute the t-value for b and construct a 99 % confidence interval for β.

68. Use the data from question 47 to compute the sample correlation coefficient r between x and y. Use a t-test to test the significance of r.
69. Suppose x and y are bivariately normally distributed. Use a t-test to test the significance of the sample correlation coefficient r if $r = .79$ and $n = 12$.
70. It is of interest to find the relationship between the monthly closing stock prices of IBM (y; unit: $10) and the closing indices of Dow Jones Industrial Average (x; unit 100).

Year/month	IMB	DJIA
2010/1	117.64	10,067.33
2010/2	122.77	10,325.26
2010/3	123.82	10,856.63
2010/4	124.54	11,008.61
2010/5	121.55	10,136.63
2010/6	119.83	9,774.02
2010/7	124.6	10,465.94
2010/8	120.08	10,014.72
2010/9	130.82	10,788.05
2010/10	140.04	11,118.40
2010/11	138.57	11,006.02
2010/12	143.76	11,577.51
2011/1	158.69	11,891.93
2011/2	159.2	12,226.34
2011/3	160.37	12,319.73
2011/4	167.75	12,810.54
2011/5	166.87	12,569.79
2011/6	169.46	12,414.34
2011/7	179.64	12,143.24
2011/8	170.56	11,613.53
2011/9	173.49	10,913.38
2011/10	183.18	11,955.01
2011/11	187.27	12,045.68
2011/12	183.17	12,217.56

Estimate the slope of the regression of the monthly closing prices of IBM against those of DJIA

71. Using the data in Problem **70**, do an F-test for the significance of the regression at $\alpha = .05$.

72. Using the data in Problem **70**, test the hypothesis that $\beta > 0$ at $\alpha = .05$.
73. Using the data in Problem **70**, construct a 95 % confidence interval for β. Is it possible for the true β to be negative?
74. Suppose we expect the DJIA for 2012/1 to be 12600. Using the data in Problem **70**, forecast the closing price for IBM and construct a 95 % prediction interval for it.

Appendix 1: Impact of Measurement Error and Proxy Error on Slope Estimates

The data collected for business and economics research are sometimes subject to errors in measurement. Recall from Chap. 2 that there are two classifications of data. Primary data are collected by the researcher specifically for a study. Secondary data are applicable to the study in question but were collected for some other reason. Survey data collected by a researcher to determine voting preference in an election are primary data. Stock prices appearing in the *Wall Street Journal* are secondary data; they were not collected for a particular study. Both primary and secondary data are subject to *measurement error,* such as computer programming errors, errors resulting from inaccurate measuring equipment, and deviations from sample statistics and population parameters. In addition to measurement error, *proxy error* can occur when a researcher uses data that do not match their theoretical definition. In other words, proxy error is the error caused by using one measurement in place of (as a proxy for) another measurement. For example, accounting income from the income statement is frequently used as a proxy for economic income to determine company value. However, accounting income is subject to changing accounting methods, and this characteristic can affect the measurement of the trends in a firm's earning power. In economics, current income (GNP) is often used as a proxy for permanent income in investigating the consumption function. Permanent income, which equals current income adjusted for transitory income, should be used instead.

If the independent variable of the regression, x_i, is subject to either measurement error or proxy error, then the observed x_i can be defined as

$$x_i^* = x_i + \eta_i, \tag{14.30}$$

where x_i is the true value of the independent variable and x_i^* is the observed value of x_i measured with errors. It is assumed that the measurement error, η_i, is independent of x_i. That is, $\text{COV}(\eta_i, x_i) = 0$ in this case, and the observed linear regression becomes

$$y_i = a + b(x_i + \eta_i) + (e_i - b\eta_i) \tag{14.31}$$

If we let $e_i - b\eta_i = e_i^*$, then Eq. 14.31 can be rewritten as

$$y_i = a + bx_i^* + e_i^* \tag{14.32}$$

In Eq. 14.32, the independent variable, x_i^*, is no longer uncorrelated with the residual e_i^*.[6]

We can illustrate Eqs. 14.30 or 14.31 by using a simple version of Friedman's (1957) theory of consumption function. In this kind of consumption function, the consumer's income is assumed to consist of a permanent component and a transitory component. The transitory component is that part of income that the consumer considers accidental. The consumption decision is determined by the permanent component. In Eq. 14.30, x_i^* represents current income, x_i represents permanent income, and $(x_i^* - x_i)$ represents transitory income. Hence, Eq. 14.30 represents a current income of the consumption function.

Equation 14.32 violates one of the assumptions of standard linear regression analysis: the error term is not independent of the observed independent variable, x_i^*. Therefore, the estimated slope, b, is no longer an unbiased estimator for β. That is,

$$E(\hat{b}) = \frac{\beta}{1 + \frac{(n-1)\sigma_\eta^2}{n\sigma_x^2}} \tag{14.33}$$

Equation 14.33 implies that if the independent variable, x_i, is measured with errors, then the ordinary least-squares estimate of β, which is b, will be a downward-biased estimate of β. When σ_η^2 approaches zero, $E(b)$ approaches β. If σ_η^2/σ_x^2 is known, then an unbiased estimate of β is

$$b' = \left(1 + \frac{n-1}{n}\frac{\sigma_\eta^2}{\sigma_x^2}\right) b \tag{14.34}$$

However, the ratio σ_η^2/σ_x^2 is seldom known. For example, in business and economics, we often use accounting income as a proxy for economic income in regression analysis. Therefore, the problem of proxy or measurement error looms large. For demonstration purposes, if $n = 10$ and $\sigma_\eta^2/\sigma_x^2 = 1/3$, then

$$E(b) = \frac{\beta}{1 + (9/10)(1/3)} = .77\beta$$

This example shows how proxy error can make the slope estimate downward-biased.

6

$$\begin{aligned}\mathrm{Cov}(x_i^*, e_i^*) &= \mathrm{Cov}[(x_i + \eta_i), (e_i - b\eta_i)] \\ &= \mathrm{Cov}(x_i, e_i) + \mathrm{Cov}(x_i, -b\eta_i) + \mathrm{Cov}(\eta_i, e_i) + \mathrm{Cov}(\eta_i, -b_{\eta i}) \\ &= -b\ \mathrm{Var}(\eta_i) \neq 0\end{aligned}$$

Appendix 2: The Relationship Between the *F*-Test and the *t*-Test

As we saw in Sect. 9.3, the t distribution with υ degrees of freedom (t_υ) can be defined as

$$t_\upsilon = \frac{Z}{\sqrt{U/\upsilon}} \tag{14.35}$$

where Z is a standard normal variable, U is a chi-square random variable with υ degrees of freedom, and Z and U are independent.

From Eq. 14.35, we can obtain

$$t_\upsilon^2 = \frac{Z^2}{U/\upsilon}$$

From Sect. 9.4, we know that Z^2 is a chi-square distribution with 1degrees of freedom. Hence t_υ^2 represents a ratio between two independent chi-square distributions. From Sect. 9.5, we know that t_υ^2 represents an F distribution with 1 and υ degrees of freedom. From this result, we can conclude that the calculated F-value should always equal the square of the calculated t_υ-value. Here $t_\upsilon - 6.0707$, so $t_\upsilon^2 = (6.0707)^2 = 36.8534$, which differs from $F = 36.8555$ only because of rounding errors.

Appendix 3: Derivation of Variance for Alternative Forecasts

Derivation of Eq. 14.17

$$\begin{aligned}
\mathrm{Var}\left(\hat{y}_{n+1}\right) &= \mathrm{Var}(\bar{y}) \\
&= \mathrm{Var}\left(\sum_{i=1}^{n} y_i/n\right) \\
&= \frac{1}{n^2}\mathrm{Var}(y_1 + \cdots + y_n) \\
&= \frac{1}{n^2}\left[\sigma_\varepsilon^2 + \cdots + \sigma_\varepsilon^2\right] \\
&= \frac{\sigma_\varepsilon^2}{n}
\end{aligned}$$

If we use the sample estimate s_e^2 for σ_ε^2, the estimate of Var (y_{n+1}) becomes

$$s^2\left(\hat{y}_{n+1}\right) = \frac{s_e^2}{n} \tag{14.36}$$

Derivation of Eq. 14.18

$$\begin{aligned}
\mathrm{Var}(\hat{y}_{n+1}) &= \mathrm{Var}[\bar{y} + b(x_i - \bar{x})] \\
&= \mathrm{Var}(\bar{y}) + \mathrm{Var}[b(x_i - \bar{x})] \\
&= \frac{\sigma_\varepsilon^2}{n} + (x_i - \bar{x})^2 \mathrm{Var(b)} \\
&= \frac{\sigma_\varepsilon^2}{n} + (x_i - \bar{x})^2 \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\end{aligned}$$

If we use the sample estimate s_e^2 for σ_ε^2, the estimate of $\mathrm{Var}(y_{n+1})$ becomes

$$s^2\left(\hat{y}_{n+1}\right) = s_e^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})} \right) \tag{14.37}$$

Derivation of Eq. 14.19

$$\mathrm{Var}(\hat{y}_{n+1} - y_{n+1,i}) = \mathrm{Var}(\hat{y}_{n+1}) + \mathrm{Var}(y_{n+1,i}) \tag{14.38}$$

The sample estimate of $\mathrm{Var}(y_{\mathrm{n+1,i}})$ can be defined as

$$\hat{\sigma}^2(y_{n+1,i}) = s_e^2 \tag{14.39}$$

Using Eqs. 14.37 and 14.39, we obtain the sample estimate of Var $(\hat{y}_{n+1} - y_{n+1,i})$ as

$$s^2\left(\hat{y}_{n+1} - y_{n+1,i}\right) = s_e^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

Chapter 15
Multiple Linear Regression

Chapter Outline

Key Terms

Multiple linear regression	Multicollinearity
Multiple regression analysis	Residual standard error
Conditional mean	Coefficient of determination
Partial regression coefficient	Mean response
Three-dimensional regression graph	Actual value
Regression plane	Individual response
Autocorrelation	Conditional prediction
Serial correlation	Cross-sectional regression
Perfect collinearity	Stepwise regression

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_15, © Springer Science+Business Media New York 2013

15.1 Introduction

Chapters 13 and 14 examined in detail the simple regression model with one independent variable (such as amount of fertilizer) and one dependent variable (such as yield of corn). In many cases, however, more than one factor can affect the outcome under study. In addition to fertilizer, rainfall and temperature certainly influence the yield of corn. In business, not only rates of return for the stock market at large affect the return on General Motors or Ford stock. Other variables, such as leverage ratio, payout ratio, and dividend yield also contribute. Therefore, regression analysis with more than one independent variable is an important analytical tool.

The model that extends a simple regression to use with two or more independent variables is called a *multiple linear regression*. Simple linear regression analysis (see Chaps. 13 and 14) helps us determine the relationship between two variables or predict the value of one variable from our knowledge of another. *Multiple regression analysis*, in contrast, is a technique for determining the relationship between a dependent variable and more than one independent variable. In addition, it can be used to employ several independent variables to predict the value of a dependent variable.

In this chapter, we first discuss the assumptions of the multiple regression model. Then we consider the method of least-squares estimation for a multiple regression model, the standard error of the residual estimate, and the coefficient of determination. Tests on sets and individual regression coefficients and forecasts in terms of a multiple regression are also investigated. Finally, we consider applications of the multiple regression model in business and economics.

15.2 The Model and Its Assumptions

In this section, we first review the simple regression model and extend it to a multiple regression model. Then we define and analyze the regression plane for two independent variables. Finally, the important assumptions we must make to use the multiple regression model are explored in some detail.

15.2.1 The Multiple Regression Model

In multiple regression, simple regression is extended by introducing more than one independent variable. Recall from Chap. 13 that a simple linear regression model can be defined as $Y_i = \alpha + \beta X_i + e_i$ and its estimate as $y_i = a + bx_i + e_i$. The sample intercept a and the sample slope b are estimates for α and β, respectively.

The normal equations used to estimate unknown parameters α and β are

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

and

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

The foregoing equations from simple linear regression are the starting point for our exploration of multiple regression in this chapter.

Suppose an individual's annual salary (Y) depends on the number of years of education (X_1) and the number of years of work experience (X_2) the individual has had. The population regression model is

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{15.1}$$

and its estimate is

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + e_i \tag{15.2}$$

where Eqs. 15.1 and 15.2 represent the multiple population regression line and the multiple sample regression line, respectively. In Eq. 15.1, α is the intercept of the regression; β_1, is the slope that represents the conditional relationship between Y and X_1, assuming X_2 is fixed; and β_2 is the slope that represents the conditional relationship between Y and X_2, assuming X_1 is fixed. If the model defined in Eq. 15.1 is linear, then the relationship between Y and each of the independent variables can be described by a straight line. In other words, the *conditional mean* of the dependent variable is given by the following population regression equation:

$$E(Y_i|X_1 = x_1, X_2 = x_2) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

The coefficients β_1 and β_2 are called *partial regression coefficients*. They indicate only the partial influence of each independent variable when the influence of all other independent variables is held constant. Just as in simple regression, the multiple sample regression line of Eq. 15.2 can be used to estimate the multiple population regression line of Eq. 15.1.

15.2.2 The Regression Plane for Two Explanatory Variables

Let us say that the stock price per share (y) can be modeled as a function of both dividend per share (x_1) and retained earnings (x_2) per share.[1]

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + e_i$$

[1] Practical examples based on Eq. 15.2 will be explored in the applications section of this chapter.

where $y_i(P_i)$ = stock price per share for the ith firm, $x_{1i}(D_i)$ = dividend per share for the ith firm, and $x_{2i}(\mathrm{RE}_i)$ = retained earnings per share for the ith firm. (Retained earnings per share equals earnings per share minus dividend per share.) The first goal of the analysis is to obtain the estimated multiple regression model

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i} \tag{15.2a}$$

The value of b_1 indicates that after the influence of the retained earnings per share is taken into account, a \$1 increase in the dividend per share (D_i) will increase the mean value of the price per share (P_i) by b_1 other things being equal. Similarly, a \$1 increase in retained earnings per share will increase the mean price per share by b_2. If there is only one explanatory variable, the estimated regression equation generates a straight line, as we saw in Chap. 13. There are two explanatory variables in Eq. 15.2a, so it represents a *regression plane (three-dimensional regression graph)*. On this three-variable regression plane, a combination of three observations (one for the value of y, one for x_1, and one for x_2) represents a single point. These points can be depicted on a three-dimensional scatter diagram. In Fig. 15.1, the best-fitted regression plane would pass near the actual sample observation points indicated by the symbol $\times$, some falling above the plane and some below in such a way as to minimize L in

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{15.3}$$

where y_i and $\hat{y}_i$ are as defined in Eqs. 15.2 and 15.2a, respectively.[2]

If there are k independent variables, then Eq. 15.1 can be generalized to

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i \tag{15.4}$$

The following section explains how regression parameters are estimated via the least-squares estimation method discussed in Chap. 13.

15.2.3 Assumptions for the Multiple Regression Model

As in simple regression analysis, we need five assumptions to perform a regression analysis of the model defined in Eq. 15.4.

1. The error term e_i is distributed with conditional mean zero and variance σ_ϵ^2 for $i = 1,2, \ldots, n$.
2. The error term e_i is independent of each of the k independent variables $x_1, x_2, \ldots, X_k$. In other words, there are no measurement errors associated with any independent variables (see Appendix 1 of Chap. 14).

[2] Using Eq. 15.3 to estimate regression parameters will be discussed in Sect. 15.3.

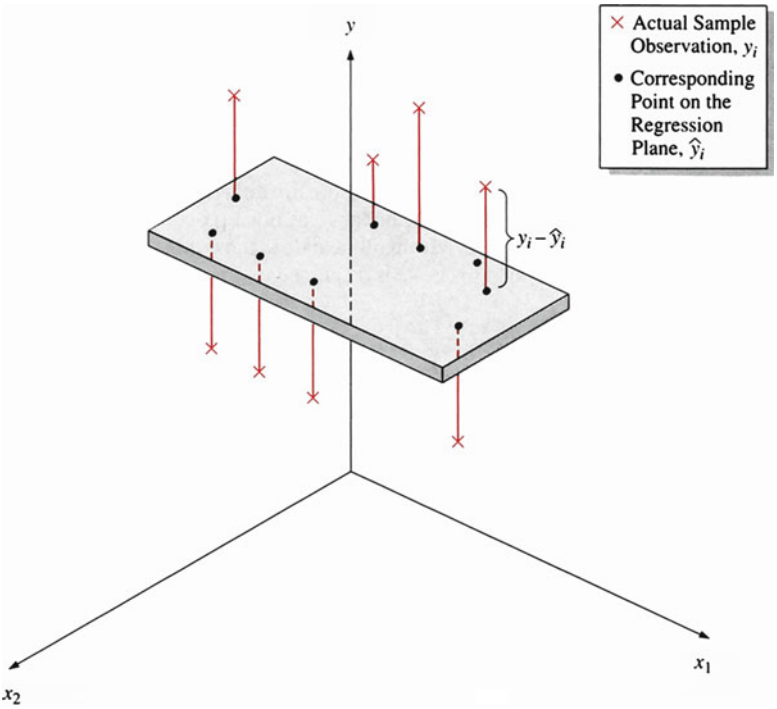


Fig. 15.1 Regression plane with $y_i(P_i)$ as dependent variable and with $x_{1i}(D_i)$ and $x_{2i}(\mathrm{RE}_i)$ as independent variables

3. Any two errors e_i and e_j are not correlated with one another; that is, their covariance is zero: $\mathrm{Cov}(e_i, e_j) = 0$ for $i \neq j$. This assumption means that there is no *autocorrelation (serial correlation)* among residual terms. This issue is discussed further in Chap. 16.
4. The independent variables are not *perfectly* related to each other in a linear function. In other words, it is not possible to find a set of numbers $d_0, d_1, d_2, \ldots, d_k$ such that

$$d_0 + d_1 X_{1i} + d_2 X_{2i} + \cdots + d_k X_{ki} = 0, \quad i = 1, 2, \ldots, n$$

In practice, the linear relationship among independent variables is usually not perfect. When a perfect linear relationship occurs, a condition known as *perfect collinearity* exists. *Multicollinearity* is the condition in which two variables are highly correlated. This issue is discussed in greater detail in Chap. 16.

15.3 Estimating Multiple Regression Parameters

To estimate the best-fitted regression plane, we use the least-squares method to estimate the regression parameters. The principle of using the least-squares method for estimating the parameters of one population regression model is demonstrated in Eq. 15.3 and Fig. 15.1. Taking Eq. 15.2 as an example, we estimate the coefficients a, b_1, and b_2 by minimizing

$$L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i})^2$$

Using the same principle and technique (Appendix 1 of Chap. 13), we can obtain the normal equations for estimating a, b_1, and b_2.[3]

$$\begin{aligned} na + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} &= \sum_{i=1}^{n} y_i \\ a \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i} x_{2i} &= \sum_{i=1}^{n} x_{1i} y_i \\ a \sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{1i} x_{2i} + b_2 \sum_{i=1}^{n} x_{2i}^2 &= \sum_{i=1}^{n} x_{2i} y_i \end{aligned} \tag{15.5}$$

If we substitute $(x_{1i} - \bar{x}_1)$, $(x_{2i} - \bar{x}_2)$ and $(y_i - \bar{y})$ for x_{1i}, x_{2i}, and y_i, then the normal equations reduce to[4]

$$\begin{aligned} b_1 \sum_{i=1}^{n} x_{1i}'^2 + b_2 \sum_{i=1}^{n} x_{1i}' x_{2i}' &= \sum_{i=1}^{n} x_{1i}' y_i' \\ b_1 \sum_{i=1}^{n} x_{1i}' x_{2i}' + b_2 \sum_{i=1}^{n} x_{2i}'^2 &= \sum_{i=1}^{n} x_{2i}' y_i' \end{aligned} \tag{15.6}$$

There are two equations and two unknowns, b_1 and b_2, associated with this equation system. Hence, we can solve b_1 and b_2 uniquely by substitution.

[3] Equation 15.5 is a three-equation simultaneous equation system with three unknowns. The values of these three unknowns can be obtained by solving this system of simultaneous equations, by using the formula derived in this section, or by using an appropriate computer package (see Sect. 15.8).

[4] In this new coordinate system, $\Sigma_{i=1}^{n} x_{1i}, \Sigma_{i=1}^{n} x_{2i}$, and $\Sigma_{i=1}^{n} y_i$ become $\Sigma_{i=i}^{n} (x_{1i} - \bar{x}_1) = 0$, $\Sigma_{i=1}^{n} (x_{2i} - \bar{x}_2) = 0$, and $\Sigma_{i=1}^{n} (y_i - \bar{y}) = 0$. If we set $x_{1i}' = x_{1i} - \bar{x}_1, x_{2i}' = x_{2i} - \bar{x}_2$ and $y_i' = y_i - \bar{y}$, then Eq. 15.5 reduce to Eq. 15.6.

$$b_1 = \frac{\left(\sum_{i=1}^{n} x'_{1i}y'_i\right)\left(\sum_{i=1}^{n} x'^2_{2i}\right) - \left(\sum_{i=1}^{n} x'_{2i}y'_i\right)\left(\sum_{i=1}^{n} x'_{1i}x'_{2i}\right)}{\left(\sum_{i=1}^{n} x'^2_{1i}\right)\left(\sum_{i=1}^{n} x'^2_{2i}\right) - \left(\sum_{i=1}^{n} x'_{1i}x'_{2i}\right)^2} \tag{15.7}$$

$$b_2 = \frac{\left(\sum_{i=1}^{n} x'^2_{1i}\right)\left(\sum_{i=1}^{n} x'_{2i}y'_t\right) - \left(\sum_{i=1}^{n} x'_{1i}x'_{2i}\right)\left(\sum_{i=1}^{n} x'_{1i}y'_i\right)}{\left(\sum_{i=1}^{n} x'^2_{1i}\right)\left(\sum_{i=1}^{n} x'^2_{2i}\right) - \left(\sum_{i=1}^{n} x'_{1i}x'_{2i}\right)^2} \tag{15.8}$$

From the estimated b_1 and b_2, we obtain the estimated regression line

$$\hat{y}'_i = b_1 x'_{1i} + b_2 x'_{2i} \tag{15.9}$$

It can be shown that the intercept of Eq. 15.2 is estimated as[5]

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \tag{15.10}$$

Example 15.1 Annual Salary, Years of Education, and Years of Work Experience. Let us use the hypothetical data given in Table 15.1 to demonstrate the procedure for estimating a multiple regression. In Table 15.1, y represents an individual's annual salary (in thousands of dollars), x_1 represents that individual's years of education, and x_2 represents her or his years of work experience.

From the data of Table 15.1, we estimate the regression line

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i} \tag{15.11}$$

The worksheet for estimating this regression line is given in Table 15.2. (This table is included to show how computers calculate mean, variance, and covariance. You do not need to remember the procedure.)

Substituting information from Table 15.2 into Eqs. 15.7, 15.8, and 15.10, we obtain

[5] Using the definitions of $\hat{y}'_i, x'_{1i},$ and $x'_{2i},$ we can rewrite Eq. 15.9 as

$$(\hat{y}_i - \bar{y}) = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2)$$

which becomes

$$\hat{y}_i = (\bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2) + b_1 x_{1i} + b_2 x_{2i} \tag{15.9′}$$

Table 15.1 Data for Example 15.1

	x_{1i}	x_{2i}	y_i
	5	7	15
	10	5	17
	9	14	26
	13	8	24
	15	*6*	*27*
Total	52	40	109.0
Mean	10.4	8	21.8

Table 15.2 Worksheet for estimating a regression line (Example 15.1)

	x_{1i}	x_{2i}	y	a	b	c	aa	bb	cc
	5	7	15	−5.4	−1	−6.8	29.16	1	46.24
	10	5	17	−.4	−3	−4.8	.16	9	23.04
	9	14	26	−1.4	6	4.2	1.96	36	17.64
	13	8	24	2.6	0	2.2	6.76	0	4.84
	15	6	27	4.6	−2	5.2	21.16	4	27.04
Mean	10.4	8	21.8						
Total	52	40	109	0	0	0	59.2	50	118.8

	$(x_{1i} - \bar{x}_1)(y_i - \bar{y})$	$(x_{2i} - \bar{x}_2)(y_i - \bar{y})$	$(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$
	ac	***bc***	***ab***
	36.72	6.8	5.4
	1.92	14.4	1.2
	−5.88	25.2	−8.4
	5.72	0	0
	23.92	−10.4	−9.2
Total	62.4	36	−11

$$\hat{b}_1 = \frac{(62.4)(50) - (36)(-11)}{(59.2)(50) - (-11)^2} = \frac{3516}{2839} = 1.2385$$

$$\hat{b}_2 = \frac{(59.2)(36) - (-11)(62.4)}{(59.2)(50) - (-11)^2} = \frac{2817.6}{2839} = .99246$$

$$\hat{a} = 21.8 - (1.2385)(10.4) - (.99246)(8) = .980$$

Hence, the regression line of Eq. 15.11 becomes

$$\hat{y}_i = .980 + 1.2385x_{1i} + .9925x_{2i} \tag{15.12}$$

The next section shows how to compute standard errors of estimates and the coefficients of determination.

15.4 The Residual Standard Error and the Coefficient of Determination

As in the case of simple regression, the standard error of estimate can be used as an absolute measure and the coefficient of determination as a relative measure of how well the multiple regression equation fits the observed data. The interpretations of these two goodness-of-fit measures are analogous to those discussed in Chap. 13.

15.4.1 The Residual Standard Error

Just like simple regression, multiple regression can be used to break down the total variation of a dependent variable y_i into unexplained variation and explained variation.

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\substack{\text{Sum of Squares}\\ \text{Total}\\ \text{(SST)}}} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\substack{\text{Sum of Squares}\\ \text{Error}\\ \text{(SSE)}}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\substack{\text{Sum of Squares}\\ \text{due to Regression}\\ \text{(SSR)}}} \tag{15.13}$$

Equation 15.13 is identical to Eq. 13.16 except that the estimated dependent variable $(\hat{y}_i)$ of multiple regression is determined by two or more independent variables. SSR and SSE are the explained and unexplained sums of squares, respectively.

Using the definition of sum of squares error, we can define the estimate of the standard deviation of error terms, sometimes called the *residual standard error*, as

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-3}} \tag{15.14}$$

Because there are three parameters—a, b_1, and b_2—for Eq. 15.2 that we must estimate before calculating the residual, the number of degrees of freedom is $(n-3)$. In other words, $(n-3)$ sample values are "free" to vary. More generally, the number of degrees of freedom for estimating the residual standard error for Eq. 15.4 is $[n-(k+1)]$.

Example 15.2 Computing y_i, e_i, and e_i^2. Using the data presented in Example 15.1, we can estimate y_i, e_i, and e_i^2 as shown in Table 15.3.

Here $\hat{y}_i$ is obtained by substituting x_{1i} and x_{2i} into Eq. 15.12. For example, $14.1198 = .980 + 1.2385(5) + .9925(7)$; $\hat{e}_i = y_i - \hat{y}_i$.

Table 15.3 Actual values, predicted values, and residuals for annual salary regression

			Residuals	
	Actual value, y_i	Predicted value, $\hat{y}_1$	e_i	e_i^2
	15	14.1198	.8802	.7748
	17	18.3272	−1.3272	1.7615
	26	26.0209	−.0209	.0004
	24	25.0200	−1.0200	1.0404
	27	25.5120	1.4880	2.2141
Total	109	–	–	5.7912

$$\sum_{i=1}^{5}(y_i-\hat{y}_i)^2=(15=14.1198)^2+(17-18.3272)^2+(26-26.0209)^2$$
$$+(24-25.0200)^2+(27-25.5120)^2=5.7912$$

Hence,

$$s_e=\sqrt{\frac{5.7912}{5-3}}=1.7016$$

s_e is one of the important components in determining the distribution of estimated a, b_1, and b_2 and fitted dependent variable ($\hat{y}$).

15.4.2 The Coefficient of Determination

We can use Eq. 15.13 to calculate a relative measure of the goodness of fit for a multiple regression.

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\text{explained variation of } y \text{ (SSR)}}{\text{total variation of } y \text{ (SST)}}=1-\frac{\text{SSE}}{\text{SST}} \tag{15.15}$$

The *coefficient of determination* R^2 is the proportion of total variation in y(SST) that is explained by the intercept and the independent variable x_1 and x_2. Note that both R^2 and S_e can be used to measure the goodness of fit for a regression. However, R^2 is a relative measure and S_e an absolute measure. Now we use the ANOVA table given in Table 15.4 to calculate the relationship between R^2 and S_e for the general multiple regression model in Eq. 15.4.

There are four columns in Table 15.4. Column (1) represents the sources of variation, column (2) alternative sums of squares that are identical to those discussed in Eq. 15.13, column (3) degrees of freedom associated with each source of variation, and column (4) the mean squares. Note that alternative mean squares represent alternative variance estimates. Mean square due to the regression is also called

Table 15.4 Notation of analysis of variance table

(1) Source of variation	(2) Sum of squares	(3) Degrees of freedom	(4) Mean square
Due to regression	$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k	SSR/k
Residual	$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n - k - 1$	SSE/$(n - k - 1)$
Total	$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n - 1$	SST/$(n - 1)$

explained variance; mean square due to the residuals is also called unexplained variance; and mean square due to the total variation can also be called variance of the dependent variable. Using those estimates, we can obtain an adjusted (or corrected) coefficient of determination $\bar{R}^2$.

$$\bar{R}^2 = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k-1} \tag{15.16}$$

The difference between R^2 and $\bar{R}^2$ is that $\bar{R}^2$ is adjusted for degrees of freedom for both SSE and SST. $\bar{R}^2$ is always smaller than R^2. If the sample size becomes large, however, $\bar{R}^2$ approaches R^2. $\bar{R}^2$ can generally help us avoid overestimating the goodness of fit for a regression relationship by adding more independent variables (relevant or not) to a regression equation. Note that the standard error of estimate (Eq. 15.14) also has been adjusted for the degrees of freedom $(n - k - 1)$.

If we divide components in Eq. 15.13 by their related degrees of freedom, then it can be shown that

$$\underbrace{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}_{\substack{\text{Total}\\ \text{Variance}}} \neq \underbrace{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-k-1}}_{\substack{\text{Unexplained}\\ \text{Variance}}} + \underbrace{\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{k}}_{\substack{\text{Explained}\\ \text{Variance}}} \tag{15.17}$$

so the adjusted coefficient of determination can be redefined as

$$\bar{R}^2 = 1 - \frac{\text{unexplained variance}}{\text{total variance}} \tag{15.16'}$$

Using the example of the last section, we can calculate the analysis of variance of Table 15.4 as shown in Table 15.5.

From Table 15.5, we can calculate R^2 and $\bar{R}^2$.

Table 15.5 Analysis of variance results

Source of variation	Sum of squares	Degrees of freedom	Mean square
Due to regression	113.0088	$k = 2$	56.5044
Residual	*5.7912*	*5 − 2 − 1 = 2*	2.8956
Total	118.8	5 − 1 = 4	29.7

$$R^2 = 113.0088/118.8 = .95125$$
$$\bar{R}^2 = 1 - \frac{2.8956}{29.7} = 1 - .09749 = .90251$$

Both R^2 and $\bar{R}^2$ imply that more than 90 % of the variation of annual salary can be explained by years of education and years of work experience. However, $\bar{R}^2$ is 4.874 % smaller than that of R^2.

15.5 Tests on Sets and Individual Regression Coefficients

After having estimated the regression model, we would like to know whether the dependent variable is related to the independent variables. To find out, we can test whether an individual regression coefficient or a set of regression coefficients is significantly different from zero. As we saw in Chap. 14, the t statistic is to test an individual coefficient and the F statistic to test linear restrictions on the parameters or regression coefficients. For this purpose, we need to assume that ε_i is normally distributed.

Logically, we perform the joint test first. If the joint test is not significant, then there is no need for the individual tests, and we normally abandon or modify the model. If the joint test is rejected, we must find out which regression coefficients are significant, so we perform individual tests.

15.5.1 Test on Sets of Regression Coefficients

Until now, our discussion has been limited to point estimation of multiple regression coefficients, the coefficient of determination, and the standard error of estimate. Now we will discuss how to use the F statistic to test whether all true population regression (slope) coefficients equal zero. The F-test rather than the t-test is used. The null hypothesis for our case is

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ &H_1 : \text{At least one } \beta \text{ is not zero.} \end{aligned} \tag{15.18}$$

If the null hypothesis is not true, then each $\hat{y}_i$ will differ from $\bar{y}$ substantially, and the explained variation $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ will be large relative to the unexplained residual variation $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. In other words, the R^2 indicated in Eq. 15.15 is relatively large. Thus, we can construct the F ratio as indicated in Eq. 15.19 to test whether the null hypothesis can be rejected.

$$F_{k,n-k-1} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/k}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-k-1)} \tag{15.19}$$

The F ratio we have constructed is the ratio of two mean square errors, as we noted in the last section, and they are two unbiased estimates of variances. Following the definition of the F distribution established in Chaps. 9 and 14, we know that the F ratio has an F distribution with k and $(n-k-1)$ degrees of freedom. This F ratio enables us to test whether at least one of the regression coefficients is significantly different from zero.

Consider the case $k = 2$. If there is no regression relationship (i.e., if $\beta_1 = \beta_2 = 0$) and because

$$\begin{aligned}\hat{y}_i &= a + b_1 x_{1i} + b_2 x_{2i} \\ &= \bar{y} + b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2)\end{aligned}$$

the $\hat{y_i}$ will be close or equal to $\bar{y}$, so the F-value will be smaller or close to zero. Thus, we cannot reject the null hypothesis that ail regression coefficients are insignificantly different from zero.

Substituting related data from Table 15.5 into Eq. 15.19, we obtain

$$\begin{aligned}F &= \frac{113.0088/2}{5.7912/2} = \frac{56.5044}{2.8956} \\ &= 19.514\end{aligned}$$

From Table A6 of Appendix A, we find that the critical value for a significance level of $\alpha = .05$ is $F_{.05,2.2} = 19.0$, which is smaller than 19.514. Therefore, we can conclude that at least one of the regression coefficients is significantly different from zero. Thus, there is a regression relationship in the population, and the improvement of explanatory power achieved by fitting a regression plane is not due to chance. In other words, the null hypothesis that years of education and years of work experience contribute nothing to an individual's annual salary is rejected at a 5 % level of significance.

Finally, the relationship between the R^2 indicated in Eq. 15.15 and the F statistic in Eq. 15.19 can be shown to be[6]

$$F_{k,n-k-1} = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2}$$

15.5.2 Hypothesis Tests for Individual Regression Coefficients

In the last section, we used the F statistic to do a joint test about a regression relationship. Now we want to use the t statistic to test whether multiple regression coefficients are significantly different from zero.

15.5.2.1 Hypothesis-Testing Specification

We follow the procedure of the last chapter to define the null hypothesis and alternative hypothesis for testing individual multiple regression coefficients.

1. Two-tailed test

$$\begin{aligned} &\mathrm{H}_0 : \beta_j = 0 \quad (j = 1,\ 2,\ \ldots, k) \\ &\mathrm{H}_1 : \beta_j \neq 0 \end{aligned} \tag{15.20}$$

2. One-tailed test

$$\begin{aligned} &\mathrm{H}_0 : \beta_j = 0 \quad (j = 1,\ 2,\ \ldots, k) \\ &\mathrm{H}_1 : \beta_j > 0 \text{ or } \beta_j < 0 \end{aligned} \tag{15.21}$$

Let's look at Eq. 15.12 as an example. For convenience, the estimated regression line is repeated here.

$$\hat{y}_i = .980 + 1.2385x_1 + .9925x_2$$

[6] Because $R^2 = 1 - \mathrm{SSE/SST} = \mathrm{SSR/SST}$,

$$\frac{R^2}{1-R^2} = \frac{\mathrm{SSR/SST}}{\mathrm{SSE/SST}} = \frac{\mathrm{SSR}}{\mathrm{SSE}}$$

In this equation, besides the estimated intercept (α) and slopes (β_1 and β_2), we have estimated the standard error of estimate for $\hat{y}_i$ as $S_e = 1.7016$. To perform the null hypothesis test, we need to know the sample distribution of b_j and the t statistic as defined in the equation

$$t_{n-k-1} = (b_j - 0)/s_{bj} \tag{15.22}$$

where $t_{n\,-\,k\,-\,1}$ represents a t statistic with $(n - k - 1)$ degrees of freedom, $k =$ the number of independent variables, and s_{bj} represents the standard error associated with b_j. The concepts and procedure used to calculate s_{bj} are similar to those used for simple regression. However, s_{bj} is quite tedious to calculate by hand; fortunately, its value is readily available in the computer output of any standard regression analysis program. Thus, in practice, we find t simply by finding the ratio of the coefficient to its estimated standard error. When the calculated value of t exceeds the critical value $t_{\alpha,n-k-1}$ indicated in the t distribution table, the null hypothesis of no significance can be rejected. We conclude that the jth independent variable x_j does have an important influence on the dependent variable y_i after the influence of all other independent variables in the model is taken into account.

15.5.2.2 Performing the t-Test for Multiple Regression Slopes

To perform the t-test for multiple regression coefficients b_1 and b_2, we estimate the sample variance of the coefficients b_1 and b_2 in accordance with Eqs. 15.23 and 15.24[7]

$$\begin{aligned}\mathrm{Var}(b_1) = s_{b_1}^2 &= \frac{s_e^2}{(1-r^2)\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2} \\ &= \frac{s_e^2\left(\sum_{i=1}^{n} x_{2i}'^2\right)}{\left(\sum_{i=1}^{n} x_{1i}'^2\right)\left(\sum_{i=1}^{n} x_{2i}'^2\right) - \left(\sum_{i=1}^{n} x_{1i}' x_{2i}'\right)^2}\end{aligned} \tag{15.23}$$

$$\begin{aligned}\mathrm{Var}(b_2) = s_{b_2}^2 &= \frac{s_e^2}{(1-r^2)\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2} \\ &= \frac{s_e^2\left(\sum_{i=1}^{n} x_{1i}'^2\right)}{\left(\sum_{i=1}^{n} x_{1i}'^2\right)\left(\sum_{i=1}^{n} x_{2i}'^2\right) - \left(\sum_{i=1}^{n} x_{1i}' x_{2i}'\right)^2}\end{aligned} \tag{15.24}$$

[7] Derivations of Eqs. 15.23 and 15.24 can be found in Appendix 1. Note that these two equations are generally estimated by computer packages (see Sect. 15.8). Manual approaches are presented here to show how sample variances of multiple regression slopes are actually calculated.

where r represents the correlation coefficient between x_{1i} and x_{2i}. If the magnitude of r is great, a collinearity problem might exist. This issue will be discussed in detail in the next chapter.

Substituting the required numerical values obtained from Tables 15.2 and 15.3, we calculate sample variances of b_1 and b_2 for Eq. 15.16.

$$\begin{aligned} S_{b_1}^2 &= \frac{(2.8956)(50)}{(59.2)(50) - (-11)^2} \\ &= \frac{(2.8956)(50)}{2839} \\ &= .05100 \end{aligned}$$

and

$$\begin{aligned} s_{b_2}^2 &= \frac{(2.8956)(59.2)}{2839} \\ &= .06038 \end{aligned}$$

Then $S_{b_1} = .2258$ and $S_{b_2} = .2457$. Dividing b_1 and b_2 by S_{b_1} and S_{b_2}, we obtain t-values for b_1 and b_2.

$$t_{b_1} = \frac{1.2385}{.2258} = 5.4849$$
$$t_{b_2} = \frac{.9925}{.2457} = 4.0395$$

Because $n = 5$ and $k = 2$, from Table A4 in Appendix, A the critical value for a one-tailed test on either coefficient (at a significance level of $\alpha = .05$) is

$$t_{\alpha,n-k-1} = t_{.05,2} = 2.920$$

We choose a one-tailed test because a priori theoretical propositions were that both x_1 and x_2 were positively related to y. Comparing 5.4849 and 4.0395 with 2.920, we conclude that both years of education and years of work experience are significantly related to an individual's annual salary.

Figure 15.2 presents all the estimates and hypothesis-testing information we have discussed in the last three sections. This example certainly proves that multiple regression analysis can be more efficiently performed by using the MINITAB computer program.

```
MTB > READ C1-C3
DATA> 5 7 15
DATA> 10 5 17
DATA> 9 14 26
DATA> 13 8 24
DATA> 15 6 27
DATA> END
      5 rows read.
MTB > REGRESS C3 2 C1 C2;
SUBC> DW;
SUBC> PREDICT 6 5.
```

Regression Analysis

```
The regression equation is
C3 = 0.98 + 1.24 C1 + 0.992 C2

Predictor       Coef       StDev          T        P
Constant       0.980       3.439       0.29    0.802
C1            1.2385      0.2258       5.48    0.032
C2            0.9925      0.2457       4.04    0.056

S = 1.702       R-Sq = 95.1%     R-Sq(adj) = 90.3%

Analysis of Variance

Source        DF          SS          MS          F        P
Regression     2     113.009      56.504      19.51    0.049
Error          2       5.791       2.896
Total          4     118.800

Source        DF      Seq SS
C1             1      65.773
C2             1      47.236

Obs         C1          C3         Fit   StDev Fit    Residual    St Resid
  1        5.0      15.000      14.120       1.499       0.880        1.09
  2       10.0      17.000      18.327       1.076      -1.327       -1.01
  3        9.0      26.000      26.021       1.632      -0.021       -0.04
  4       13.0      24.000      25.020       0.961      -1.020       -0.73
  5       15.0      27.000      25.512       1.301       1.488        1.36


Durbin-Watson statistic = 2.39
```

Fig. 15.2 MINITAB output of multiple regression in terms of data given in Table 15.1

15.6 Confidence Interval for the Mean Response and Prediction Interval for the Individual Response

15.6.1 Point Estimates of the Mean and the Individual Responses

One of the important uses of the multiple regression line is to obtain predictions and forecasts for the dependent variable, given an assumed set of values of the independent variables. This kind of prediction is called the *conditional prediction* (forecast), just as in simple regression (see Sect. 14.4, of which this model is an extension). Suppose the independent variables are equal to some specified values $x_{1,n+1}$ and $x_{2,n+1}$, and that the linear relationship among y_n, $x_{1,n}$, and $x_{2,n}$ continues to hold.[8] Then the corresponding value of the dependent variable y_{n+1} is

$$Y_{n+1,i} = \alpha + \beta_1 x_{1,n+1,i} + \beta_2 x_{2,n+1,i} + \varepsilon_{n+1,i} \tag{15.25}$$

which, given $x_{1,n+1}$ and $x_{2,n+1}$, has expectation

$$E(Y_{n+1}|x_{1,n+1,}x_{2,n+1}) = \alpha + \beta_1 x_{1,n+1} + \beta_2 x_{2,n+1} \tag{15.26}$$

Equation 15.26 yields the *mean response* $E(Y_{n+1}|x_{1,n+1}, x_{2,n+1})$ that we want to estimate when the independent variables are fixed at $x_{1,n+1}$ and $x_{1,n+1}$. Equation 15.25 yields the *actual value* (or *individual response*) that we want to predict.

To obtain the best point estimate, we first estimate the sample regression line as defined in Eq. 15.2. Then we substitute the given values $x_{1,n+1}$ and $x_{2,n+1}$ into the estimated Eq. 15.12, obtaining

$$\hat{y}_{n+1} = a + b_1 x_{1,n+1} + b_2 x_{2,n+1} \tag{15.27}$$

This is the best point estimate for both conditional expectation and actual-value forecasts. In other words, the forecast of conditional expectation value is equal to the forecast of actual value. However, the forecasts are interpreted differently. The importance of these different interpretations will emerge when we investigate the process of making interval estimates.

15.6.2 Interval Estimates of Forecasts

To construct a confidence interval for forecasts, it is necessary to know the distribution, mean, and variance of $\hat{y}_{n+1}$. The distribution of $\hat{y}_{n+1}$ is a *normal* distribution. The variance associated with $\hat{y}_{n+1}$ may be classified into three cases.

[8] $x_{1,n+1}$ and $x_{2,n+1}$ can be either given values or forecasted values. When a regression is used to describe a time-series relationship, they are forecasted values.

First, we deal with a case in which the conditional mean ($\hat{y}_{n+1}$) is equal to the unconditional mean ($\bar{y}$). In the second and third cases, we deal with the conditional mean. However, case 2 involves the mean response and case 3 the individual response.

Case 15.1 Conditional Expectation (Mean Response) with $x_{1,n+1} = \bar{x}_1$ and $\bar{x}_{2,n+1} = \bar{x}_2$

From the definitions of the intercept of a regression or the sample regression line, we have

$$\begin{aligned}\hat{y}_{n+1} &= (\bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2) + b_1x_{1,n+1} + b_2x_{2,n+1} \\ &= \bar{y} + b_1(x_{1,n+1} - \bar{x}_1) + b_2(x_{2,n+1} - \bar{x}_2)\end{aligned}$$

If $x_{1,n} = \bar{x}_1$ and $x_{2,n} = \bar{x}_2$, then $\hat{y}_{n+1} = \bar{y}$. Following Appendix 3 of Chap. 14, we obtain the estimate of the variance for y_{n+1} as

$$s^2(\hat{y}_{n+1}) = s^2(\bar{y}) = s_e^2/n \tag{15.28}$$

Case 15.2 Conditional Expectation (Mean Response) with $x_{1,n+1} \neq \bar{x}_1$ or $x_{2,n+1} \neq \bar{x}_2$

In this case, the forecast value can be defined as

$$\hat{y}_{n+1} = \bar{y} + b_1(x_{1,n+1} - \bar{x}_1) + b_2(x_{2,n+1} - \bar{x}_2) \tag{15.29}$$

Following Appendix 2, we obtain the estimate of the variance for $\hat{y}_{n+1}$ in terms of sample standard variance of estimates S_e^2 as

$$\begin{aligned}s_1^2 &= s^2(\hat{y}_{n+1}) \\ &= s_e^2\left[\frac{1}{n} + \frac{(x_{1,n+1} - \bar{x}_1)^2}{(1-r^2)C_1^2} + \frac{(x_{2,n+1} - \bar{x}_2)^2}{(1-r^2)C_2^2} - \frac{2(x_{1,n+1} - \bar{x}_1)(x_{2,n+1} - \bar{x}_2)r}{(1-r^2)C_1C_2}\right]\end{aligned} \tag{15.30}$$

where $C_1 = \sqrt{\Sigma_{i=1}^n (x_{1,j} - \bar{x}_1)^2}$, $C_2 = \sqrt{\Sigma_{i=1}^n (x_{2,i} - \bar{x}_2)^2}$ and r = correlation coefficient between $x_{1,i}$ and $x_{2,i}$.

Case 15.3 Actual Value (Individual Response) of y_{n+1}

After we have derived the sample variance for $\hat{y}_{n+1}$, we derive the sample variance for individual response (observation), $y_{n+1,i}$ (which deviates from $\hat{y}_{n+1}$ by a random error e_i).

$$y_{n+1,i} = \hat{y}_{n+1} + e_i$$

The variance of an individual observation, $y_{n+1,i}$, includes the variance of the observation about the regression line $\left(s_e^2\right)$ as well as $s^2(\hat{y}_{n+1,i})$. Because $\hat{y}_{n+1}$ and e_i are independent, $s^2(y_{n+1,j}) = s^2(\hat{y}_{n+1}) + s_e^2$. More explicitly,

$$s^2(\hat{y}_{n+1,i}) = s_1^2 + s_e^2 = s_2^2 \tag{15.31}$$

where s_1^2 is defined in Eq. 15.30

Using Eqs. 15.28, 15.30, and 15.31, we can obtain a confidence interval for prediction as follows:

1. For prediction of the conditional expectation with $x_{1,n+1} = \bar{x}_1$ and $x_{2,n+1} = \bar{x}_2$, the confidence interval is

$$\hat{y}_{n+1} \pm t_{n-3,\alpha/2} \frac{s_e}{\sqrt{n}} \tag{15.32}$$

2. For prediction of the conditional expectation with $x_{1,n+1} \neq \bar{x}_2$ or $x_{1,n+1} \neq \bar{x}_2$, the confidence interval is

$$\hat{y}_{n+1} \pm (t_{n-3,\alpha/2}) s_1 \tag{15.33}$$

where s_1, is defined in Eq. 15.30.

3. For prediction of the actual value $y_{n+1,i}$ the prediction interval is

$$\hat{y}_{n+1,i} \pm (t_{n-3,\alpha/2}) s_2 \tag{15.34}$$

where s_2 is defined in Eq. 15.31.

To show how Eq. 15.34 is applied in constructing the confidence interval for forecasting the actual value of y_{n+1}, let's use the annual salary example (Table 15.2) to find the 95 % prediction interval for annual salary, y_{n+1}, when a person has 6 years of education and 5 years of work experience. The predicted annual salary can be computed from Eq. 15.12.

$$\begin{aligned} \hat{y}_{n+1,i} &= .980 + (1.2385)(6) + (.9925)(5) \\ &= 13.3735 \text{ (in thousands of dollars)} \end{aligned}$$

From Table 15.2, we have

$$\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 = 59.2, \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 = 50, \; n = 5$$

$$\sum_{i=1}^{n} (x_{1i} - \bar{x}_i)(x_{2i} - \bar{x}_2) = -11, \; \bar{x}_1 = 10.4, \; \bar{x}_2 = 8$$

Using this information, we calculate

$$C_1 = \sqrt{59.2} = 7.6942, \qquad C_2 = \sqrt{50} = 7.0711$$

$$r = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}} = \frac{-11}{(7.6942)(7.0711)}$$

$$= -.2022, \; r^2 = .0409$$

$$(x_{1,n+1} - \bar{x}_1)^2 = (6 - 10.4)^2 = 19.36$$

$$(x_{2,n+1} - \bar{x}_2)^2 = (5 - 8)^2 = 9$$

From Table 15.3, we have $S_e^2 = 5.7912/(5-3) = 2.8956$. Substituting this information into Eq. 15.31 yields

$$s_2^2 = (2.8956)\left[1 + \frac{1}{5} + \frac{19.36}{(1 - .0409)(59.2)} + \frac{9}{(1 - .0409)(50)} - \frac{2(-.2022)(-4.4)(-3)}{(1 - .0409)(7.6942)(7.0711)}\right] = (2.8956)(1.83) = 5.2989$$

We will use $n = 5, s_2 = \sqrt{5.2989} = 2.3019$, and $\hat{y}_{n+1,i} = 13.3735$. From Table A4, in Appendix A, we have $t_{0.025,2} = 4.303$. Substituting this information into Eq. 15.34, we find that the annual salary is predicted with 95 % confidence by the interval

$$13.3735 \pm (4.303)(2.3019) = 13.3735 \pm 9.9051$$

$$3.4684 \leq y_{n+1,i} \leq 23.2786$$

When n is large, we can modify this expression by replacing t with the appropriate normal deviate z.

MINITAB output showing prediction results of $x_{1,n+1,i} = 6$ and $x_{2,n+1,i} = 5$ is presented in Fig. 15.3. The prediction interval shown in the last row of Fig. 15.3 is (3.466, 23.280), which is similar to what we calculated before.

In the next two sections, we will explore applications of multiple regression in business and economics. Section 15.8 explicitly treats the use of SAS and MINITAB computer programs to do multiple regression analyses.

15.7 Business and Economic Applications

Multiple regression analysis has been widely used in decision making in business and economics. Five examples are discussed in this section.

```
MTB > READ C1-C3
DATA> 5 7 15
DATA> 10 5 17
DATA> 9 14 26
DATA> 13 8 24
DATA> 15 6 27
DATA> END
      5 rows read.
MTB > BRIEF 3
MTB > REGRESS C3 2 C1 C2;
SUBC> DW.
```

Regression Analysis

```
The regression equation is
C3 = 0.98 + 1.24 C1 + 0.992 C2

Predictor       Coef        StDev           T        P
Constant       0.980        3.439        0.29    0.802
C1            1.2385       0.2258        5.48    0.032
C2            0.9925       0.2457        4.04    0.056

S = 1.702       R-Sq = 95.1%     R-Sq(adj) = 90.3%

Analysis of Variance

Source        DF           SS           MS          F         P
Regression     2      113.009       56.504      19.51     0.049
Error          2        5.791        2.896
Total          4      118.800

Source       DF      Seq SS
C1            1      65.773
C2            1      47.236

Obs         C1          C3         Fit  StDev Fit   Residual    St Resid
  1        5.0      15.000      14.120      1.499      0.880        1.09
  2       10.0      17.000      18.327      1.076     -1.327       -1.01
  3        9.0      26.000      26.021      1.632     -0.021       -0.04
  4       13.0      24.000      25.020      0.961     -1.020       -0.73
  5       15.0      27.000      25.512      1.301      1.488        1.36


Durbin-Watson statistic = 2.39

     Fit  StDev Fit        95.0% CI              95.0% PI
  13.373      1.551    (   6.699,  20.047)  (   3.466,  23.280)
```

Fig. 15.3 MINITAB output of $y_{n+1,i}$

Application 15.1 Overall Job-Worth of Performance for Certain Army Jobs. Bobko and Donnelly (1988) employed multiple regression to estimate overall job-worth to the army of certain army jobs from attributes of those jobs.[9] Their final regression prediction model is

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + b_4x_{4i} + b_5x_{5i} + b_6x_{6i} + b_7x_{7i} + e_i$$

where

y_i = job-value judgments of overall worth for the ith individual
x_{1i} = performance level for the ith job
x_{2i} = combat probability for the ith job
x_{3i} = enlistment bonus for the ith job
x_{4i} = reenlistment bonus for the ith job
x_{5i} = aptitude required for entry into the ith job
x_{6i} = cost of error for the ith job
x_{7i} = job variety for the ith job

Bobko and Donnelly estimated this multiple regression model using data obtained from interviews. Their regression results are presented in Table 15.6. As would be expected, performance level was the single best predictor of 95 estimates of judgments of overall worth. The other job-level correlates were combat probability, enlistment bonus, reenlistment bonus, aptitude, cost of error, and task variety. The first six of these predictors had statistically significant regression weights (coefficients), p-value $< .05$, indicating their unique contribution to the prediction of overall worth estimates. However, task variety was not statistically significant.

Application 15.2 The Relationship Between Individual Stock Rates of Return, Payout Ratio, and Market Rates of Return. To demonstrate multiple regression analysis, a time-series regression for 1970–2009 is run, the dependent variable being the rate of return for the JNJ stock ($R_{j,t}$) and the independent variables being the payout ratio (dividend per share/earnings per share) for JNJ ($P_{j,j}$) and the rates of return on the S&P 500 Index, $R_{m,t}$. The results are as follows:

$$R_{j,t} = \alpha_j + \gamma_j P_{j,t} + \beta_j R_{m,t} + \epsilon_{j,t}$$

Fortunately, the results do not have to be calculated by hand but can be obtained by using MINITAB. The MINITAB results are presented in Table 15.7. The parameter value for the market rates of return is 0.7329, which is called the beta coefficient. A 1 % increase in the market rate of return will lead to a 0.7329 % change in the rate of return of the JNJ stock, given the payout ratio—that is, the rate of return of JNJ stock is less volatile than that of the market. The payout ratio has a

[9] P. Bobko and L. Donnelly (1988), "Identifying Correlations of Job-Level, Overall Worth Estimates: Application in a Public Sector Organization," *Human Performance* 3, 187–204

Table 15.6 Best subset regression of overall worth on job-level predictors

Source	df	Sum of squares	*F*	*P*
Regression	7	16.007	274.12	.0001
Error (residual)	87	.726		
Variable	*Regression weight*		*t-ratio*	*p*
Performance level	.013		1666.22	.0001
Combat probability	.039		21.19	.0001
Enlistment bonus	.034		18.52	.0001
Reenlistment bonus	.016		15.73	.0001
Aptitude	.013		26.01	.0001
Cost of error	.029		5.61	.0201
Task variety	.016		2.51	.1166

Source: Bobko and Donnelly (1988), *Human Performance*
Note: Adjusted $R^2 = .953$; $n = 95$ mean estimates of overall worth

Table 15.7 $R_{j,t} = \alpha_j + \gamma_j P_{j,t} + \beta_j R_{m,t} + \varepsilon_{j,t}$

Variable	Coefficient	Standard error	t-value	*p*-value
Constant	0.0777	0.2049	0.3793	0.7066
Payout ratio	−0.2133	0.5613	−0.3800	0.7061
Market rate of return	0.7329	0.3510	2.0880	0.0437
$R^2 = 0.106$				
$\bar{R}^2 = 0.057$				
F-value = 2.184				
Observations 40				

coefficient of −0.2133. This result implies that a 1 % increase in the payout ratio will lead to a 0.2133 % decrease in the mean rate of return on JNJ stock, given the market rate of return.

The independent variables are statistically significant at the 5 % level. The *t*-value for the market is 2.0880, and the associated *p*-value is 0.0437, which means that the lowest level of significance at which the null hypothesis can be rejected is 4.37 %. This suggests that the population coefficient for the market is not equal to zero. The *t* statistic for the payout ratio, which is calculated by dividing the parameter value (−0.2133) by the standard error (0.5613), is −0.3800. Its *p*-value is 0.7061, thus the null hypothesis cannot be rejected.

R^2 for the regression is 0.106. In other words, the independent variables explain about 10.6 % of the variation in the rate of return on JNJ stock. The adjusted *R*-square, $\bar{R}^2$, which takes into account overfitting in the sample, is equal to 0.057.

The *F*-value, which tests the hypothesis that the population coefficients of the independent variables are both zero against the alternative that they are not, is equal to 2.184. The degrees of freedom associated with this *F*-value are $v_1 = 2$ and $v_2 = 37$. From Table A6 in Appendix A, we find that the critical value for the *F*-test is $F_{.01,2,30} = 5.39$ and $F_{.01,2,40} = 5.18$. Because the *F*-value for the regression is less than the critical value 5.39, the null hypothesis cannot be rejected.

Application 15.3 Analyzing the Determination of Price per Share. To further demonstrate multiple regression techniques, let us say that a *cross-sectional regression* is run. In a cross-sectional regression, all data come from a single period. The dependent variable in this regression is the price per share (P_j) of the 30 firms used to compile the Dow Jones Industrial Average for the year 2009. The independent variables are the dividend per share (DPS_j) and the retained earnings per share (EPS_j) for the 30 firms. (Retained earnings per share is defined as earnings per share minus dividend per share. Price per share is the close price of the end of year 2009; dividend per share and retained earnings per share are based on 2009 annual balance sheet and income statement.) The sample regression relationship is

$$P_j = a + b_1 DPS_j + b_2 \mathrm{EPS}_j + e_j \quad (j = 1,\ 2, \ldots, 30)$$

Empirical results are presented in Table 15.8. The constant term is significant with a t-value of 2.518. This result means that the intercept term is statistically different from zero and the null hypothesis can be rejected at both a10 and a 5 % level. The retained earnings per share variable is highly significant with a t-value of 4.478 and a p-value of 0.000. Thus, we can reject the null hypothesis that the coefficient is equal to zero and accept the alternative hypothesis that it differs from zero and makes a contribution to price per share. The coefficient for this variable is 0.978; mean price per share increases $0.978 when the retained earnings per share increases by $1.00, given the dividend.

The coefficient for the dividend per share variable has a t-value of 2.756 and a p-value of .010. This is the lowest level of significance at which the null hypothesis can be rejected; thus the null hypothesis is rejected at both a 10 % and a 5 % level. The coefficient for dividend per share is 12.836. When the dividend increases by $1.00, the price per share tends to rise by $12.836.

The value of R^2 is 0.724, which means that the model explains 72.4 % of the observed fluctuations in the price per share. The adjusted R-square, $\bar{R}^2$, is 0.703. The F-value for the regression is 35.35. The number of degrees of freedom for the regression and residual are 2 and 27, respectively. The critical value for F at a 1 % level of significance is 5.49. Because the regression F-value is greater than the critical value, the null hypothesis that the coefficients are equal to zero is rejected.

Application 15.4 Multiple Regression Approach to Evaluating Real Estate Property. To show how the multiple regression technique can be used by real estate appraisers, Andrews and Ferguson (1986) used the data in Table 15.9 to do the multiple regression analysis.

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i$$

where

y_i = sale price for ith house
x_{1i} = home size for ith house
x_{2i} = condition rating for ith house

Table 15.8 $P_j = a + b_1 DPS_j + b_2\, EPS_j + e_j$

Variable	Coefficient	Standard error	t-value	p-value
Constant	12.800	5.084	2.518	0.018
DPS	12.836	4.657	2.756	0.010
EPS	0.978	0.218	4.478	0.000
$R^2 = 0.724$				
$\bar{R}^2 = 0.703$				
F-value = 35.35				
Observations 30				

Table 15.9 Sale price, house size, and condition rating

Sale price, y (thousands of dollars)	Home size, x_1 (hundreds of sq. ft.)	Condition rating, x_2 (1 to 10)
60.0	23	5
32.7	11	2
57.7	20	9
45.5	17	3
47.0	15	8
55.3	21	4
64.5	24	7
42.6	13	6
54.5	19	7
57.5	25	2

Source: R. L. Andrews and J. T. Ferguson, "Integrating Judgment with a Regression Appraisal." *The Real Estate Appraiser and Analyst*, Vol. 52, No. 2, Spring 1986 (Table 1)

MINITAB regression outputs in terms of Table 15.9 are presented in Fig. 15.4. From this output, the estimated regression is

$$\begin{aligned} \hat{y}_i &= 9.782 + 1.87094x_{1i} + 1.2781x_{2i} \\ &\quad (6.00) \quad (24.56) \quad (8.85) \end{aligned}$$

t-values are in parenthesis.

From Table A4 in Appendix A, we find that $t_{.025,7} = 2.365$. Because t-values for 3 regression parameters are larger than 2.365, all estimated parameters are significantly different from 0 at $\alpha = .05$. This estimated regression can be used to estimate the sale price for a house. For example, if $x_1 = 18$ and $x_2 = 5$, the predicted sale price is

$$\begin{aligned} \hat{y}_i &= 9.781 + (1.87094)(18) + (1.2781)(5) \\ &= 49.8484 \end{aligned}$$

```
MTB > READ C1-C3
DATA> 60.0 23 5
DATA> 32.7 11 2
DATA> 57.7 20 9
DATA> 45.5 17 3
DATA> 47.0 15 8
DATA> 55.3 21 4
DATA> 64.5 24 7
DATA> 42.6 13 6
DATA> 54.5 19 7
DATA> 57.5 25 2
DATA> END
     10 rows read.
MTB > BRIEF 1
MTB > REGRESS C1 2 C2 C3;
SUBC> DW.
```

Regression Analysis

```
The regression equation is
C1 = 9.78 + 1.87 C2 + 1.28 C3

Predictor       Coef        StDev           T         P
Constant       9.782        1.630        6.00     0.000
C2           1.87094      0.07617       24.56     0.000
C3            1.2781       0.1444        8.85     0.000

S = 1.081       R-Sq = 99.0%     R-Sq(adj) = 98.7%

Analysis of Variance

Source       DF          SS          MS          F         P
Regression    2      819.33      409.66     350.87     0.000
Error         7        8.17        1.17
Total         9      827.50


Durbin-Watson statistic = 1.56

MTB > CORRELATION C1-C3
```

Correlations (Pearson)

```
             C1       C2
C2        0.938
C3        0.373    0.043
```

Fig. 15.4 MINITAB output for Table 15.9

This implies that the estimated sale price is $49,848.4 if the home size is 18,000 square feet and the condition rating is 5.

Application 15.5 Multiple Regression Approach to Doing Cost Analysis. To show how the multiple regression technique can be used to do cost analysis by accountants, we look at Benston's research. Benston (1966) used a set of sample data (as shown in Table 15.10) from a firm's accounting and production records to provide cost information about the firm's shipping department to do the multiple regression analysis

$$y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + b_3 x_{3t} + e_t$$

where

y_t = hours of labor in tth week
x_{1t} = thousands of pounds shipped in tth week
x_{2t} = percentage of units shipped by truck in tth week
x_{3t} = average number of pounds per shipment in tth week

MINITAB regression output is presented in Fig. 15.5. From p-values indicated in Fig. 15.5, we find that b_0 and b_3 are significantly different from 0 at $\alpha = .01$. Hence, we can conclude that the only important variable in determining the hours of labor required in the shipping department is the average number of pounds per shipment.

15.8 Using Computer Programs to Do Multiple Regression Analyses

15.8.1 SAS Program for Multiple Regression Analysis

In an example taken from Churchill's *Marketing Research*, data for the sales of Click ballpoint pens (y), advertising (x_1, measured in TV spots per month), number of sales representatives (x_2), and a wholesaler efficiency index (x_3) were presented in Table 14.10 of the last chapter.

In Sect. 14.6, we investigated only the relationship between two variables (y and x_1, y and x_2, and y and x_3). Now we will expand that analysis by using the following three regression models:[10]

$$y_i = a + b_1 x_{1i} + e_i \tag{15.a}$$

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + e_i \tag{15.b}$$

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + e_t \tag{15.c}$$

[10] In these regressions, we hold the price of a ballpoint pen and the income of a consumer constant, because this is a set of cross-sectional data.

Table 15.10 Hours of labor and related factors cause costs to be incurred

Week	Hours of labor, y	Thousands of pounds shipped, x_1	Percentage of units shipped by truck, x_2	Average number of pounds per shipment, x_3
1	100	5.1	90	20
2	85	3.8	99	22
3	108	5.3	58	19
4	116	7.5	16	15
5	92	4.5	54	20
6	63	3.3	42	26
7	79	5.3	12	25
8	101	5.9	32	21
9	88	4.0	56	24
10	71	4.2	64	29
11	122	6.8	78	10
12	85	3.9	90	30
13	50	3.8	74	28
14	114	7.5	89	14
15	104	4.5	90	21
16	111	6.0	40	20
17	110	8.1	55	16
18	100	2.9	64	19
19	82	4.0	35	23
20	85	4.8	58	25

Source: G. J. Benston (1966), "Multiple Regression Analysis of Cost Behavior," *Accounting Review*, Vol. 41, No. 4, 657–672 (Reprinted by permission of the publisher)

Equation 15.a can be used to investigate the relationship between y and x_1, which was discussed in Sect. 14.6.

Equation 15.b can be used to analyze whether the second explanatory variable, x_2, improves the equation's power to explain the variation of sales. Equation 15.c can be used to analyze whether the third explanatory variable, x_3, further improves that explanatory power. Part of the output of the SAS program for Eqs. 15.a, 15.b, and 15.c is presented in Fig. 15.6a–c. Figure 15.6a shows the regression results of Eq. 15.a, Fig. 15.6b the regression results of Eq. 15.b, and Fig. 15.6c the regression results of Eq. 15.c. Using these results, we will review and summarize simple regression and multiple regression results that have been discussed in Chaps. 13, 14, and 15.

Computer outputs of Fig. 15.6a–c present the following results of simple and multiple regression.

1. Estimated intercept and slopes
2. F-values for the whole regression
3. t-values for individual regression coefficients
4. ANOVA of regression
5. $R2$ and $\bar{R}^2$
6. p-values

```
MTB > READ C1-C4
DATA> 100 5.1 90 20
DATA> 85 3.8 99 22
DATA> 108 5.3 58 19
DATA> 116 7.5 16 15
DATA> 92 4.5 54 20
DATA> 63 3.3 42 26
DATA> 79 5.3 12 25
DATA> 101 5.9 32 21
DATA> 88 4.0 56 24
DATA> 71 4.2 64 29
DATA> 122 6.8 78 10
DATA> 85 3.9 90 30
DATA> 50 3.8 74  28
DATA> 114 7.5 89 14
DATA> 104 4.5 90 21
DATA> 111 6.0 40 20
DATA> 110 8.1 55 16
DATA> 100 2.9 64 19
DATA> 82 4.0 35 23
DATA> 85 4.8  58 25
DATA> END
     20 rows read.
MTB > REGRESS C1 3 C2 C3 C4;
SUBC> DW.
```

Regression Analysis

```
The regression equation is
C1 = 132 + 2.73 C2 + 0.0472 C3 - 2.59 C4

Predictor       Coef        StDev           T         P
Constant      131.92        25.69        5.13     0.000
C2             2.726        2.275        1.20     0.248
C3           0.04722      0.09335        0.51     0.620
C4           -2.5874       0.6428       -4.03     0.001

S = 9.810        R-Sq = 77.0%     R-Sq(adj) = 72.7%

Analysis of Variance

Source        DF          SS          MS          F         P
Regression     3      5158.3      1719.4      17.87     0.000
Error         16      1539.9        96.2
Total         19      6698.2

Durbin-Watson statistic = 2.43
```

Fig. 15.5 MINITAB output for Application 15.5

a

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	463451.00888	463451.00888	130.644	0.0001
Error	38	134802.01487	3547.42144		
C Total	39	598253.02375			

Root MSE	59.56023	R-square	0.7747
Dep Mean	411.28750	Adj R-sq	0.7687
C.V.	14.48141		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	135.433596	25.90650568	5.228	0.0001
X1	1	25.307698	2.21415038	11.430	0.0001

Durbin-Watson D	1.721
(For Number of Obs.)	40
1st Order Autocorrelation	0.133

b

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	522778.45899	261389.22949	128.141	0.0001
Error	37	75474.56476	2039.85310		
C Total	39	598253.02375			

Root MSE	45.16473	R-square	0.8738
Dep Mean	411.28750	Adj R-sq	0.8670
C.V.	10.98130		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	69.328469	23.15546229	2.994	0.0049
X1	1	14.156185	2.66360071	5.315	0.0001
X2	1	37.531322	6.95929855	5.393	0.0001

Durbin-Watson D	2.125
(For Number of Obs.)	40
1st Order Autocorrelation	-0.083

Fig. 15.6 (continued)

c

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	527209.08074	175736.36025	89.051	0.0001
Error	36	71043.94301	1973.44286		
C Total	39	598253.02375			

Root MSE	44.42345	R-square	0.8812
Dep Mean	411.28750	Adj R-sq	0.8714
C.V.	10.80107		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	31.150390	34.17504533	0.911	0.3681
X1	1	12.968162	2.73723213	4.738	0.0001
X2	1	41.245624	7.28010741	5.666	0.0001
X3	1	11.524255	7.69117684	1.498	0.1428

Durbin-Watson D	2.104
(For Number of Obs.)	40
1st Order Autocorrelation	-0.083

Fig. 15.6 (**a**) SAS output for regression results of $y_i = a + b_1x_{1i} + e_i$. (**b**) SAS output for regression results of $y_i = a + b_1x_{1i} + b_2x_{2i} + e_i$. (**c**) SAS output for regression results of $y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + e_i$

7. Durbin–Watson D and first-order autocorrelation (these two statistics are discussed in Sect. 16.4 of Chap. 16)
8. Standard error of residual estimate (mean square of error)
9. Root MSE $= \sqrt{\text{MSE error}}$. For example, for Eq. 15.b, root MSE $= \sqrt{2039.85310}$ $= 45.16473$. The root MSE estimate can be used to measure the performance of prediction. We will explore this in further detail in Chap. 18 when we discuss time-series analysis.

These SAS regression outputs give us almost all the sample statistics we have examined so far. Now let's consider the practical implications of Eqs. 15.a, 15.b, and 15.c. In Sect. 14.6 of the last chapter, we discussed the estimated regression of Eq. 15.a.

Equation 15.b specifies a regression model in which sales are the dependent variable and the independent variables are number of TV spots x_1 and number of sales representatives x_2. The fitted regression equation is

$$\underset{}{\hat{y}} \underset{}{=} \underset{(2.994)}{69.3} \underset{}{+} \underset{(5.315)}{14.2x_1} \underset{}{+} \underset{(5.393)}{37.5x_2} \quad F = 128.141$$

Here t-values are indicated in parentheses.

This regression indicates that when the number of TV spots increases by 1 unit, sales increase by \$14.200 on average while the number of sales representatives stays unchanged. When the number of sales representatives increases by 1 person, sales increase by \$37,500 on average while the number of TV spots stays unchanged.

The F-value for the regression of Eq. 15.b is 128.141. There are 40 observations and two independent variables, so the number of degrees of freedom in the model is $40 - 2 - 1 = 37$. By interpolation, it can be shown that the critical value of $F_{.05,2,37}$ is 3.25 (Table A6 in Appendix A). Because the F-value for the regression is greater than the critical value, the hypothesis that the coefficients are equal to zero is rejected. From the t-values associated with estimated regression coefficients, we find that the estimated intercept and slopes are significant at $\alpha = .01$.

Because the t-values of b_2 are significantly different from zero, we conclude that adding the number of sales representatives improves the equation's power to explain sales. This conclusion can also be drawn from the fact that $\bar{R}^2$ has increased from .7687 to .8670.

The fitted regression of Eq. 15.c is

$$\hat{y} = \underset{(.911)}{31.1504} + \underset{5(4.738)}{12.9682x_1} + \underset{(5.666)}{41.2456x_2} + \underset{(1.498)}{11.5243x_3} \quad F = 89.051$$

Again, t-values are indicated in parentheses.

Following Sect. 15.5, we first test the whole set of regression coefficients in terms of the F statistic. From Table A6 in Appendix A, by interpolation, we find that $F_{01.3,36} = 2.88$. $F = 89.051$ is much larger than 2.88. This implies that we reject the following null hypothesis of our joint test:

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

Now we can use t statistics to test which individual coefficient is significantly different from zero. From Table A4 in Appendix A, by interpolation, we find that the critical value of t statistic is $t_{.005,36} = 2.72$. By comparing this critical value with 4.738, 5.666, and 1.498, we conclude that b_1 and b_2 are significantly different from zero and that b_3 is not significantly different from zero at $\alpha = .01$. In other words, the wholesaler efficiency index does not increase the explanatory power of Eq. 15.c.

15.8.2 MINITAB Program for Multiple Regression Prediction

MINITAB is used to run the regression defined in Fig. 15.6c and presented in Fig. 15.7. Besides regression parameters, we also predict y by assuming $x_1 = 13$, $x_2 = 9$, and $x_3 = 5$. The results are listed in the last row of Fig. 15.7. They are

1. $\hat{y_n}$+1, $_i = 628.57$
2. $s(\hat{y}_{n+1}) = 34.92$
3. $s(\hat{y}_{n+1,i}) = \sqrt{s^2(\hat{y}_{n+1} + s_e^2)} = \sqrt{(34.92)^2 + 1973} = 56.50$

4. 95 % confidence interval: (557.73, 699.40)
5. 95 % prediction interval: (513.94, 743.19)

15.8.3 Stepwise Regression Analysis

In this example, we want to use *stepwise regression* to establish a statistical model to predict the sales of Click ballpoint pens (y). We are considering three possible explanatory variables: advertising (x_1) measured in TV spots per month, the number of sales representatives (x_2), and a wholesaler efficiency index (x_3). The question is what variables should be included in the statistical model to explain the sales. The stepwise regression method suggests the following steps.

Step 1:

Run simple regression on each explanatory variable, and choose the model that explains the highest amount of variation in y. The regression results obtained are presented in Fig. 14.14a, b. The R^2-value in each computer report is used to determine which variable enters the model first. Upon comparing R^2-values for the three models, we conclude that x_2, which has the highest R^2-value (.7775), should enter the model first.

Independent variable	R^2	F-value
x_1	.7747	130.644
x_2	.7775	132.811
x_3	.0000	.000

Step 2:

The second variable to enter should be the variable that, in conjunction with the first variable, explains the greatest amount of variation in y.

Independent variables		R^2	F-value
x_2	x_1	.8738	128.141
x_2	x_3	.807	77.46

The R^2-values and F-values in the foregoing table are obtained from Figs. 15.6b to 15.8. The table shows the results when x_1 and x_3 are combined with x_2 to explain the variation in y. The combination of x_1 and x_2 clearly yields a higher R^2 (.8738). This suggests that x_1 should be the second variable to enter.

Step 3:

In this step, we want to decide whether another variable should enter the model to explain y. Note that every time an additional variable is included in a model, R^2 increases. The question is whether the increase in R^2 justifies inclusion of the variable. We apply an F-test to answer this question.

$$F = \frac{(R_f^2 - R_R^2)(k_f - k_R)}{(1 - R_f^2)/(N - k_f - 1)}$$

```
MTB > PRINT C1-C4

Data Display

 Row     C1    C2    C3    C4

   1  260.3     5     3     4
   2  286.1     7     5     2
   3  279.4     6     3     3
   4  410.8     9     4     4
   5  438.2    12     6     1
   6  315.3     8     3     4
   7  565.1    11     7     3
   8  570.0    16     8     2
   9  426.1    13     4     3
  10  315.0     7     3     4
  11  403.6    10     6     1
  12  220.5     4     4     1
  13  343.6     9     4     3
  14  644.6    17     8     4
  15  520.4    19     7     2
  16  329.5     9     3     2
  17  426.0    11     6     4
  18  343.2     8     3     3
  19  450.4    13     5     4
  20  421.8    14     5     2
  21  245.6     7     4     4
  22  503.3    16     6     3
  23  375.7     9     5     3
  24  265.5     5     3     3
  25  620.6    18     6     4
  26  450.5    18     5     3
  27  270.1     5     3     2
  28  368.0     7     6     2
  29  556.1    12     7     1
  30  570.0    13     6     4
  31  318.5     8     4     3
  32  250.2     6     3     2
  33  667.0    16     8     2
  34  618.3    19     8     2
  35  525.3    17     7     4
  36  332.2    10     4     3
  37  393.2    12     5     3
  38  283.5     8     3     3
  39  376.2    10     5     4
  40  481.8    12     5     2

MTB > NAME C1'Y' C2'X1' C3'X2' C4'X3'
MTB > BRIEF 3
MTB > REGRESS C1 3 C2 C3 C4;
SUBC> DW;
SUBC> PREDICT 13 9 5.
```

Fig. 15.7 (continued)

```
The regression equation is
Y = 31.2 + 13.0 X1 + 41.2 X2 + 11.5 X3

Predictor       Coef       StDev          T        P
Constant       31.15       34.18       0.91    0.368
X1            12.968       2.737       4.74    0.000
X2            41.246       7.280       5.67    0.000
X3            11.524       7.691       1.50    0.143

S = 44.42       R-Sq = 88.1%     R-Sq(adj) = 87.1%

Analysis of Variance

Source       DF          SS          MS          F        P
Regression    3      527209      175736      89.05    0.000
Error        36       71044        1973
Total        39      598253

Source       DF      Seq SS
X1            1      463451
X2            1       59327
X3            1        4431

Obs         X1          Y        Fit  StDev Fit    Residual    St Resid
  1        5.0     260.30     265.83      14.96       -5.53       -0.13
  2        7.0     286.10     351.20      12.82      -65.10       -1.53
  3        6.0     279.40     267.27      11.35       12.13        0.28
  4        9.0     410.80     358.94      11.54       51.86        1.21
  5       12.0     438.20     445.77      15.11       -7.57       -0.18
  6        8.0     315.30     304.73      13.18       10.57        0.25
  7       11.0     565.10     497.09      16.43       68.01        1.65
  8       16.0     570.00     591.65      15.24      -21.65       -0.52
  9       13.0     426.10     399.29      13.88       26.81        0.64
 10        7.0     315.00     291.76      13.24       23.24        0.55
 11       10.0     403.60     419.83      15.63      -16.23       -0.39
 12        4.0     220.50     259.53      18.78      -39.03       -0.97
 13        9.0     343.60     347.42       8.26       -3.82       -0.09
 14       17.0     644.60     627.67      18.77       16.93        0.42
 15       19.0     520.40     589.31      17.23      -68.91       -1.68
 16        9.0     329.50     294.65      15.87       34.85        0.84
 17       11.0     426.00     467.37      14.98      -41.37       -0.99
 18        8.0     343.20     293.21      11.61       49.99        1.17
 19       13.0     450.40     452.06      11.57       -1.66       -0.04
 20       14.0     421.80     441.98      13.88      -20.18       -0.48
 21        7.0     245.60     333.01      13.61      -87.41       -2.07R
 22       16.0     503.30     520.69      11.52      -17.39       -0.41
```

Fig. 15.7 (continued)

```
 23       9.0      375.70      388.66       9.07     -12.96      -0.30
 24       5.0      265.50      254.30      12.18      11.20       0.26
 25      18.0      620.60      558.15      16.71      62.45       1.52
 26      18.0      450.50      505.38      20.34     -54.88      -1.39
 27       5.0      270.10      242.78      13.82      27.32       0.65
 28       7.0      368.00      392.45      17.60     -24.45      -0.60
 29      12.0      556.10      487.01      16.82      69.09       1.68
 30      13.0      570.00      493.31      12.85      76.69       1.80
 31       8.0      318.50      334.45       8.63     -15.95      -0.37
 32       6.0      260.20      255.74      13.55       4.46       0.11
 33      16.0      667.00      591.65      15.24      75.35       1.81
 34      19.0      618.30      630.56      16.53     -12.26      -0.30
 35      17.0      525.30      586.43      15.37     -61.13      -1.47
 36      10.0      332.20      360.39       8.77     -28.19      -0.65
 37      12.0      393.20      427.57       7.61     -34.37      -0.79
 38       8.0      283.50      293.21      11.61      -9.71      -0.23
 39      10.0      376.20      413.16      12.25     -36.96      -0.87
 40      12.0      481.80      416.04      10.48      65.76       1.52

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 2.10

     Fit  StDev Fit        95.0% CI              95.0% PI
  628.57      34.92    (  557.73,  699.40)   (  513.94,  743.19) XX
X  denotes a row with X values away from the center
XX denotes a row with very extreme X values
```

Fig. 15.7 MINITAB output of $y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + e_i$

where

$R_f^2 = R^2$ of the model with the new variable
$R_R^2 = R^2$ of the model without the new variable
$k_f =$ number of the variables in the model with the new variable
$k_R =$ number of the variables in the model without the new variable

To determine whether x_3 should be included in the model, we need to compare the R^2 of the model with x_3 and R^2 of the model without x_3.

Independent variables			R^2
x_2	x_1		.8738
x_2	x_1	x_3	.8812

Using the foregoing formula, we compute

$$F = \frac{(.8812 - .8738)/(3 - 2)}{(1 - .8812)/(40 - 3 - 1)} = 2.24 < F_{.05,1,36} = 4.11$$

```
MTB > BRIEF 2
MTB > REGRESS C1 2 C3 C4;
SUBC> DW.
```

Regression Analysis

```
The regression equation is
Y = 5.2 + 68.7 X2 + 22.1 X3

Predictor       Coef        StDev           T         P
Constant        5.19        42.40        0.12     0.903
X2            68.744        5.523       12.45     0.000
X3            22.079        9.252        2.39     0.022

S = 55.83       R-Sq = 80.7%     R-Sq(adj) = 79.7%

Analysis of Variance

Source        DF          SS          MS          F         P
Regression     2      482914      241457      77.46     0.000
Error         37      115339        3117
Total         39      598253

Source        DF      Seq SS
X2             1      465161
X3             1       17753

Unusual Observations
Obs        X2           Y         Fit  StDev Fit    Residual    St Resid
 21      4.00      245.60      368.49      14.28      -122.89      -2.28R
 25      6.00      620.60      505.97      15.79       114.63       2.14R

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 2.22
```

Fig. 15.8 MINITAB output of $y_i = a + b_1x_2 + b_3x_3 + e_i$

Because including x_3 does not increase R^2 significantly, the null hypothesis that x_3 should not be included is not rejected in this case. Our conclusion from the stepwise regression analysis is that the best model should include only x_1 and x_2 as explanatory variables.

Some computer packages are programmed to perform the whole complicated stepwise regression in response to one simple command. Figure 15.9 shows the output of a stepwise regression analysis using MINITAB.

15.9 Summary

In this chapter, we examined multiple regression analysis, which describes the relationship between a dependent variable and two or more independent variables. Methods of estimating multiple regression (slope) coefficients and their standard errors were discussed in depth. The residual standard error and the coefficient of determination were also explored in some detail.

```
MTB > STEPWISE REGRESSION C1 3 C2 C3 C4
```

Stepwise Regression

```
 F-to-Enter:       4.00    F-to-Remove:       4.00

 Response is     Y     on  3 predictors, with N =    40

    Step         1         2
Constant     80.07     69.33

X2            66.2      37.5
T-Value      11.52      5.39



X1                      14.2
T-Value                 5.31

S             59.2      45.2
R-Sq         77.75     87.38
 More? (Yes, No, Subcommand, or Help)
SUBC> YES

 No variables entered or removed

 More? (Yes, No, Subcommand, or Help)
SUBC> NO
```

Fig. 15.9 Stepwise regression analysis

Both *t*-tests and *F*-tests for testing regression relationships were discussed in this chapter. We investigated the confidence interval for the mean response and the prediction interval for the individual response. And finally, we saw how multiple regression analyses can be used in business and economics decision making.

Questions and Problems

1. Compare simple regression to multiple regression. When would you use simple regression? When would you use multiple regression?
2. In simple regression, the geometric interpretation is to fit a line that best describes the relationship between x and y. What is the geometric interpretation of multiple regression when there are two independent variables?

3. Discuss the differences between the assumptions of the simple and the multiple linear regression models.
4. We can test the significance of a simple regression either by using a t-test to test the slope coefficient or by using an F-test to test the significance of the model. How does our approach differ when we are testing the significance of a multiple regression?
5. Explain how the number of degrees of freedom available for estimating σ^2 of the error term is related to the number of variables in the regression.
6. Briefly compare the concepts of simple correlation, partial correlation, and multiple correlation.
7. Compare the ways the regression coefficients are interpreted in simple regression and in multiple regression.
8. What is a partial regression coefficient? How do we measure it?
9. What is multicollinearity? Why is it a problem in multiple regression?
10. Suppose an NFL scout is interested in what physical attributes make for a good quarterback. He collects data on the height and weight of 8 quarterbacks and their performance ratings for the year. The data are summarized in the following table.

Performance rating, y	Height (inches), x_1	Weight (pounds), x_2
94.3	73	210
83.3	69	185
92.3	77	225
72.4	75	215
69.5	71	190
65.8	70	180
101.2	76	212
77.4	73	195

Use the MINITAB program to answer the following questions.

(a) Estimate the regression coefficients α, β_1, and β_2 and interpret the results.
(b) Compute the t-values for the coefficients and test the significance of β_1 and β_2.

11. Using the information given in question 10, compute SSR, SSE, SST, and R^2. Also use an F-test to test the significance of the model.
12. Using the results from question 10, forecast the performance rating for a quarterback who is 6 ft 1 in. tall and weighs 200 lb. Construct a 95 % confidence interval around this forecast.
13. The chairperson of the finance department at Rutgers University would like to find the relationship between undergraduate grade point average (UGPA) and GMAT scores on graduate grade point average (GGPA). She collects the following data on six students.

UGPA, x_1	GMAT, x_2	GGPA, y
3.45	485	3.62
3.10	500	3.75
3.00	525	3.81
2.95	560	3.88
3.11	575	3.85
2.87	625	3.95

(a) Calculate the regression parameters for α, β_1, and β_2.
(b) Compute the standard errors of the regression coefficients. Use a t-test to test the significance of b_1 and b_2.

14. Suppose we were interested in testing the joint significance of b_1 and b_2 in terms of data from question 13. That is, the null hypothesis is H_0: $\beta_1 = \beta_2 = 0$.

(a) Explain how we would conduct such a test.
(b) Test the joint significance of β_1 and β_2.

15. Use the data and results from question 13 to construct 90 % confidence intervals for b_1 and b_2.
16. Suppose a student has a 3.85 undergraduate GPA and a GMAT score of 575.

(a) Forecast this student's graduate GPA.
(b) Construct a 90 % confidence interval for this forecast. Use Eqs. 15.30 and 15.33.

17. Suppose a labor economist is interested in the effect of experience and education on income. He obtains the following regression.

$$\hat{\text{INCOME}} = 24,000 + 1,000(\text{EXPER}) + 500(\text{EDUC})$$

where

INCOME = income measured in dollars
EXPER = years of experience
EDUC = years of education

Interpret the regression coefficients for EXPER and EDUC.

18. Suppose you calculate $s_{b_1} = 325$ and $s_{b_2} = 285$, and you know that 50 observations were used to estimate the model. Test the significance of the regression coefficients in question 17.
19. An agent for Decade 100 Real Estate Company is interested in developing a model that explains the value of a piece of real estate. She collects data on the following variables:

Number of bedrooms	Sales price
Number of bathrooms	Age of house
Miles from main highway	Size of lot

(a) Which variables should be the independent variables?
(b) Write down a multiple regression equation that might be of interest to this realtor, and explain to her what signs to expect for the regression coefficients and how to interpret the regression coefficients.
(c) Explain the usefulness of this model.
(d) Will employing confidence intervals for forecasted values be useful in this analysis? Explain.

20. Suppose you estimate a regression using 20 observations and 16 independent variables. You compute R^2 to be .98. Explain why R^2 may not be an appropriate measure of the goodness of fit. Can you think of a better one?
21. Suppose a travel consultant is interested in the relationship between people's incomes and the amount of money they spend for vacations. He chooses to estimate the regression

$$E(\text{VAC}) = \alpha + \beta_1(\text{WSAL}) + \beta_2(\text{MSAL})$$

where

VAC = dollars spent on vacation
WSAL = weekly salary
MSAL = monthly salary

Do you think he will encounter any difficulties in estimating this model?

22. Thomas Chen, an education professor, is interested in the relationship among final exam scores, midterm exam scores, and hours studied for the final. He collects the following data.

Final exam score, y	Midterm exam score, x_1	Hours studied, x_2
75	74	5
83	89	8
72	65	9
88	92	4
95	90	10

(a) Estimate the regression coefficients for α, β_1, and β_2.
(b) Compute the estimated R^2 and of the adjusted R^2.

23. Using the data and your results from question 22, test the individual significance of β_1 and β_2. Also construct a 99 % confidence interval for β_1 and β_2.

24. Using the data and your results from question 22, forecast the final exam score for a student who scored 97 on the midterm and studied $6\frac{1}{2}$ h for the final. Construct a 90 % confidence interval for this forecast.
25. In multiple regression, we can test the significance of the individual regression coefficients by using a t-test, or we can test the joint significance of the coefficients by using an F-test. Is it possible for the t-tests to be significant while the F-test is insignificant? Explain.
26. An economist at the National Academy of Movie Theater Owners wants to estimate the demand for movie tickets. He chooses to estimate the equation.

$$\mathrm{QT}_t = \alpha + \beta_1 \mathrm{PT}_t + \beta_2(\mathrm{GNP}_t) + \varepsilon_t$$

where

QT_t = quantity of movie tickets purchased in year t
PT_t = average price of movie tickets in year t
GNP_t = gross national product in year t (in billions of dollars)

What signs do you expect for the coefficients on price and GNP to have?
Use MINITAB to answer questions 27–33.

27. Suppose the economist of question 26 collects the following data.

Year	QT	PT	GNP
1986	1,000	\$7.00	1,000
1987	1,100	7.25	1,250
1988	1,200	6.75	1,175
1989	1,300	6.50	1,800
1990	1,400	6.50	2,000
1991	1,500	6.25	2,250

(a) Estimate the demand for movie tickets.
(b) Do the coefficients carry the correct signs?
(c) If you were going to use a t-test to test the significance of b_1 and b_2, should you use a one-tailed or a two-tailed test?
(d) Use a t-test to test the significance of b_1 and b_2.

28. Use your results from question 27 to compute R^2 and $\bar{R}^2$. Also use an F-test to test the joint significance of the regression.
29. Construct 95 % confidence intervals for the coefficients b_1 and b_2 from the regression in question 27.
30. Suppose you have obtained the following 1992 and 1993 forecasts of GNP and ticket prices.

Year	GNP (in billions of dollars)	Prices
1992	2,572	7.25
1993	3,000	8.00

(a) Forecast the quantity of tickets sold for 1992 and 1993.
(b) Construct 90 % confidence intervals for these forecasts. What information do these confidence intervals provide?

31. Suppose the economist in question 26 is interested in estimating the price and income elasticity of demand for movie tickets. He can do this by taking the natural logarithms of QT, PT, and GNP and reestimating the multiple regression. Using the data from question 27, estimate the price and income elasticity for movie tickets and interpret your results.
32. Use a *t*-test to test the significance of the estimated elasticities.
33. Use an *F*-test to test the joint significance of the price and income elasticity. Use MINITAB to answer questions 34–37.
34. An investment analyst is interested in developing an equation to forecast the earnings per share of a company. He collects the following data for five companies.

Company	EPS	Sales in $	Advertising expense in $	Cost in $
1	1.00	100	80	50
2	2.00	175	120	28
3	1.50	89	72	30
4	3.00	225	175	20
5	3.25	300	240	25

(a) Formulate a suitable regression model to explain EPS.
(b) Are there any variables you may want to omit from the regression? If so, why?

35. Suppose the analyst of question 34 decides on the following regression:

$$\text{EPS} = \alpha + \beta_1(\text{SALES}) + \beta_2(\text{COST}) + \varepsilon$$

(a) Estimate the intercept and slope coefficients.
(b) Use an *F*-test to test the joint significance of the slope coefficients.
(c) Compute the standard error for a, b_1, and b_2, and use a *t*-test to test their significance.

36. Construct 90 % confidence intervals for α, β_1, and β_2, using the results from question 35.

37. Forecast the EPS for a company with $400 in sales and a cost of $65. Construct a 99 % confidence interval for this forecast. Use Eqs. 15.30 and 15.33.
38. You estimate a regression using a computer package that generates the following output.

Coefficient	Estimate	Standard error
Intercept	12.53	6.54
X_1	−9.37	5.25
X_2	14.75	4.36
X_3	.27	.09

(a) Compute the t-values for the coefficients.
(b) Say the sample used to estimate the regression consisted of 27 observations. Are the coefficients significant?

39. Use the foregoing information to construct 95 % confidence intervals for the parameters.
40. Buford Lightfoot, a stock market analyst, is interested in finding a model to describe the returns for different stocks. He estimates the following regression:

$$R_{it} = \alpha + \beta_1 R_{m,t} + \beta_2 I_{i,t} + \varepsilon_t$$

where

$R_{m,t}$ = return on the S&P 500 in month t
$R_{i,t}$ = return on stock i in month t
$I_{i,t}$ = index for stock i's industry in month t

The results of this regression are

Coefficient	Estimate	Standard error
Intercept	−3.45	2.32
R_m	1.32	.65
I_i	−.32	.10

Interpret the results of the regression and compute the t-values for the coefficients.

41. Construct a 90 % confidence interval for the parameter estimates from question 40. Assume $n = 30$.
42. Say you know that the return on the S&P 500 will be 3 % next month and that the industry index next month will be 2. Forecast stock i's return.
43. Suppose you fit the model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$$

using 35 data points and obtain SSE = .56 and $R^2 = .85$. Test the null hypothesis that all βs are equal to zero against the alternative hypothesis that at least one of the βs is nonzero. Conduct this test at the 95 % confidence level.

44. Again consider question 43. Examine SSE and R^2, and explain whether the model provides a good fit.
45. Suppose you estimate the model

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

using 25 observations and obtain

$$\Sigma(z_i - \hat{z}_i)^2 = 2.45 \quad \text{and} \quad \Sigma(z_i - \bar{z})^2 = 3.65$$

Compute R^2. Does the model provide a good fit?

46. Explain why, given the same independent variables, the confidence interval for the mean value of y is always narrower than the corresponding confidence interval for any other value of y.
47. You are given the following information:

Cov $(y, x_1) = 65$,	Mean of $y = 80$
Var $(x_1) = 4.5$	$\Sigma(x_1 - x_1)^2 = 40$
Var $(x_2) = 4.2$	$\Sigma(x_2 - x_2)^2 = 35$
Cov $(x_1, x_2) = 3.2$	$Se = 48$
Var $(y) = 2{,}500$	$n = 30$
Cov $(y, x_2) = 3.8$	Mean of $x_1 = 7$
	Mean of $x_2 = 6$

(a) Compute the least-squares estimate for intercept and slopes
(b) Compute the t-values for slopes b_1 and b_2

48. Use the data from question 47 to compute

(a) R^2
(b) Adjusted R^2
(c) F statistics

49. Use the data of question 47 and the answers of questions 47 and 48 to predict y if $x_1 = 7.5$ and $x_2 = 6.5$. In addition, please also calculate 95% confidence interval and 95 % prediction interval.
50. The admissions officer at Poindexter U. would like to determine the effect of high school GPA and SAT scores on undergraduate GPA. He collects the following data on six students.

HSGPA, x_1	SAT, x_2	UGPA, y
2.6	585	2.02
2.9	525	2.75
3.0	475	3.10
2.8	620	2.95
3.1	525	3.25
3.87	650	3.95

(a) Calculate the regression parameters for α, β_1, and β_2.
(b) Compute the standard errors of the regression coefficients. Use a t-test to test the significance of b_1 and b_2.

51. Suppose we are interested in testing the joint significance of b_1 and b_2. That is, the null hypothesis is H_0: $\beta_1 = \beta_2 = 0$. Test the joint significance of β_1 and β_1.
52. Use the data and results from question 50 to construct 90 % confidence intervals for β_1 and β_2.
53. Suppose a student with a 3.85 high school GPA and an SAT score of 555 applies for admission to Poindexter U.

    (a) Forecast this student's undergraduate GPA.
    (b) Construct a 90 % confidence interval for this forecast.

54. You have been hired as an economist for the Federal Reserve Bank of New York. Your job is to forecast future interest rates. Summarize the theory of the interest rate, and formulate a mathematical model that can be used to forecast the interest rate.
55. You have been hired as a consultant for AT&T. Formulate a mathematical model that could be used to estimate the demand for telephone service.
56. You have been hired as an economist for the Public Utility Commission of Wisconsin. The agency needs an estimate of the demand for electricity in order to determine what rates the electric companies can charge. Formulate a mathematical model that can be used to estimate the demand for electricity.
57. Researchers interested in determining the relationship between a firm's annual sales and its expenditures on research and development (x_1), television advertising (x_2), and all other advertising (x_3) run a regression analysis of 23 firms in the same industry. The results are

$$\hat{y} = -2.3 + 5.8x_1 + 4.2x_2 + 7.4x_3$$
$$\quad\quad (1.20) \quad\quad (1.31) \quad (1.56)$$

The quantities in parentheses are the standard errors of the net regression coefficients. The standard error of estimate $S_{y.123}$ is 124. The standard deviation of the dependent variable S_y is 325.

(a) Interpret the net regression coefficient b_1.
(b) Test, at the 1 % level of significance, whether each of the net regression coefficients is significantly different from zero.
(c) What is the expected effect when highly correlated independent variables are included in a multiple regression equation?
(d) Calculate the coefficient of multiple correlation and the coefficient of multiple determination.
(e) Estimate the average annual sales for a firm that has research and development expenditures of $6 million, television advertising expenditures of $10 million, and all other advertising expenditures of $7 million.

58. A statistician is interested in using product price and the amount of advertising to predict the sales of a product. Several combinations of price and advertising are tried, with the following results. (Sales are given in ten thousands of dollars, advertising in thousands of dollars.)

Sales, y (ten thousands of dollars)	12	8	9	14	6	11	10	8
Price, x_1	4	4	5	5	6	6	7	7
Advertising, x_2 (thousands of dollars)	3	0	5	7	3	8	6	8

Determine the estimated multiple regression line and the value of r^2.

59. Use the following equation to answer parts (a) through (d).

$$\hat{y} = -1.67 + 2.46x_1 - 5.48x_2$$

(a) What is the meaning of the numbers -1.67, 2.46, and -5.48?
(b) Graph the relationship between $\hat{y}$ and x_1 for $x_2 = 10$.
(c) Graph the relationship between $\hat{y}$ and x_1 for $x_2 = 20$.
(d) What is the difference between the lines you graphed in parts (b) and (c)?

60. An economist states that wages should be inversely related to the rate of unemployment and should be positively related to prices. Test these claims at the .10 Type I error level, using the following data in a multiple regression specification. Report a p-value (the significance) for each coefficient.

Average dollar wage	266	255	235	220	207	189	175
Unemployment rate	9.7	7.6	7.1	5.8	6.1	7.1	7.7
Price index	289	272	246	217	195	181	170
Average dollar wage	163	154	145	136	127	120	
Unemployment rate	8.5	5.6	4.9	5.6	5.9	4.9	
Price index	161	147	133	125	121	116	

61. How are simple and multiple regression similar? How are they different?
62. The closing prices of six stocks on the last day of last month seem to be quite highly correlated with their latest reported earnings per share figures and with the percentage of earnings growth they experienced in the past year. The figures are given in the accompanying table. Use Eq. 13.6 or a computer to show that the regression equation is $\hat{y} = .72 + 5.94x_1 + 1.08x_2$.

Closing price per share, y	Latest earnings per share, x_1	Percentage earnings growth, x_2
$10	$1.50	3
18	2.00	10
22	2.00	8
30	2.50	6
30	3.00	10
40	7.00	−1

63. A regression equation was found to be $\hat{y} = 1.5 + .2x_1 + 3.1x_2$. R^2 for this equation was .95. The values of x_1 ranged from −10 to 15 and those of x_2 from −20 to −50. Which of the following statements are true?

    (a) Variable x_2 is more strongly correlated with y than is x_1 because its regression coefficient is larger.
    (b) When $x_1 = 0$ and $x_2 = 0$, the predicted value of y is 1.5, and this value has legitimate physical meaning because the data values for x_1 and x_2 span zero.
    (c) A very high proportion of the squared error of prediction incurred by using $\bar{y}$ as the predictor can be eliminated by using the regression in making predictions.

64. A regression equation was found to be $\hat{y} = 10 + 14x_1 - 7x_2$. Which of the following statements are true?

    (a) A 1-unit increase in x_1 causes y to increase by 14 units.
    (b) Variable y is more highly correlated with x_1 than with x_2 because the coefficient of x_1 is positive.
    (c) If the value of x_2 is large enough, negative predictions of y will be obtained.

65. A regression analysis has two independent variables (x_1 and x_2).

    (a) What does it mean if x_1 and x_2 are independent of each other? In that case, what is the correlation between them?
    (b) Is saying that x_1 and x_2 are independent variables the same as saying that they are independent of each other? Explain.

Use the return information for 3-month T-bills, the NYSE Index, Chrysler, Ford, and GM for the 3-year period from January 1985 through December 1987 on Chap. 18 (pages 960–962). Suppose of interest now is to establish the relationship between NYSE Index and the other four variables based on the result of the stepwise regression method. Problems 66 to 69 are based on the following output from MINITAB.

Stepwise Regression:
NYSE Index versus T-Bill, Chrysler, Ford, GM
Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15
Response is NYSE Index on 4 predictors, with N = 36

Step	1	2	3	4
Constant	−0.0029835	−0.0003357	0.0028444	−0.0498685
Ford	0.526	0.307	0.216	0.243
T-value	9.10	3.43	2.03	2.29
P-value	0.000	0.002	0.051	0.029
Chrysler		0.207	0.178	0.150
T-value		3.02	2.54	2.11
P-value	0.005	0.016	0.043	
GM			0.16	0.18
T-value			1.51	1.66
P-value			0.140	0.107
T-bill				10.1
T-value				1.48
P-value				0.149
S	0.0328	0.0295	0.0289	0.0284
R-Sq	70.89	77.17	78.70	80.10
R-Sq(adj)	70.03	75.79	76.70	77.54
Mallows Cp	13.4	5.6	5.2	5.0

66. (a) Which explanatory variable enters the model first? What is the corresponding R^2?
    (b) Which explanatory variable enters the model last? What is the corresponding C?
67. Write down the estimated regression equation and the adjusted R^2 from the results of Step 3 and 4, respectively.
68. At $\alpha = .05$, do an F-test for the significance of the regression at Step 3.
69. At $\alpha = .05$, do an F-test for the significance of the regression at Step 2.
70. For the return data with NYSE Index being explained by Ford and Chrysler, suppose we expect the returns for Ford and Chrysler are 0.0229 and 0.0337, respectively. Forecast the return for NYSE and construct a 95 % prediction interval for it.

Appendix 1: Derivation of the Sampling Variance of the Least-Squares Slope Estimations

Using the definition of the simple correlation coefficient given in Eq. 13.24 in Chap. 13, we can obtain the correlation coefficient between x_1 an x_2 as

$$r = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{C_1 C_2} \tag{15.35}$$

where

$$C_1 = \sqrt{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2} \tag{15.36a}$$

$$C_2 = \sqrt{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2} \tag{15.36b}$$

Substituting 15A.1, 15A.2a, and 15A.2b into Eq. 15.7 yields

$$\begin{aligned} b_1 &= \frac{C_2^2\left[\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i')\right] - rC_1C_2\left[\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)(y_i')\right]}{C_1^2C_2^2 - r^2C_1^2C_2^2} \\ &= \sum_{i=1}^{n}\left[\frac{(x_{1i} - \bar{x}_1)(y_i') - (rC_1/C_2)(x_{2i} - \bar{x}_2)(y_i')}{(1 - r^2)C_1^2}\right] \\ &= \sum_{i=1}^{n}\left[\frac{(x_{1i} - \bar{x}_1) - (rC_1/C_2)(x_{2i} - \bar{x}_2)}{(1 - r^2)C_1^2}\right]y_i' \end{aligned} \tag{15.37}$$

Substituting

$$\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)y_i' = \sum_{i=1}^{n}(x_{1i} - \bar{x}_i)y_i$$

and

$$\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)y_i' = \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)y_i$$

into Eq. 15.37, and letting the coefficient of y_i equal B_{1i}, we obtain

$$b_1 = \sum_{i=1}^{n} B_{1i}y_i \tag{15.38}$$

Similarly,

$$\begin{aligned} b_2 &= \sum_{i=1}^{n}\left(\frac{(x_{2i} - \bar{x}_2) - \frac{rC_2}{C_1}(x_{1i} - \bar{x}_i)}{(1 - r^2)C_2^2}\right)y_i \\ &= \sum_{i=1}^{n} B_{2i}y_i \end{aligned} \tag{15.39}$$

Substituting Eq. 15.2 into Eqs. 15.38 and 15.39, we get

$$b_1 = a\sum_{i=1}^{n} B_{1i} + b_1\sum_{i=1}^{n} B_{1i}x_{1i} + b_2\sum_{i=1}^{n} B_{1i}x_{2i} + \sum_{i=1}^{n} B_{1i}e_i$$

and

$$b_2 = a\sum_{i=1}^{n} B_{2i} + b_1\sum_{i=1}^{n} B_{2i}x_{1i} + b_2\sum_{i=1}^{n} B_{2i}x_{2i} + \sum_{i=1}^{n} B_{2i}e_i$$

It can easily be shown that $\Sigma_{i=1}^{n} B_{1i} = 0$, $\Sigma_{i=1}^{n} B_{2i} = 0$, $\Sigma_{i=1}^{n} B_{1i}x_{1i} = 1$, $\Sigma_{i=1}^{n} B_{2i}x_{2i} = 1$, $\Sigma_{i=1}^{n} B_{2i}x_{1i} = 0$, and $\Sigma_{i=1}^{n} B_{1i}x_{2i} = 0$. Therefore, these two equations imply that

$$b_1 - E(b_1) = b_1 - \beta_1 = \sum_{i=1}^{n} B_{1i}e_i \tag{15.40}$$

and

$$b_2 - E(b_2) = b_2 - \beta_2 = \sum_{i=1}^{n} B_{2i}e_i \tag{15.41}$$

From Eq. 15.40, we obtain

$$\begin{aligned}\mathrm{Var}(b_1) &= E\left[\left(\sum_{i=1}^{n} B_{1i}e_i\right)^2\right] - \left[E\left(\sum_{i=1}^{n} B_{1i}e_i\right)\right]^2 \\ &= \sum_{j=1}^{n}\sum_{j=1}^{n} B_{1i}B_{1j}E(e_ie_j) - \left(\sum_{i=1}^{n} B_{1i}\right)^2 [E(e_i)]^2 = s_e^2\sum_{i=1}^{n} B_{1i}^2\end{aligned} \tag{15.42}$$

In Eq. 15.8, the last equality holds because $E(e_i) = 0$ and $E(e_ie_j) = 0$ when $i \neq j$. From the definition of B_{1i} in Eq. 15.38, we have

$$B_{1i}^2 = \frac{(x_{1i} - \bar{x}_1)^2 + r^2C_1^2/C_2^2(x_{2i} - \bar{x}_2)^2 - 2r(C_1/C_2)(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{(1 - r^2)^2C_1^4} \tag{15.43}$$

And from Eqs. 15.43, 15.35, 15.36a, and 15.36b,

$$\sum_{i=1}^{n} B_{1i}^2 = \frac{C_i^2 + r^2\left(C_1^2/C_2^2\right)\left(C_2^2\right) - 2r(C^1/C_2)(r_1C_1C_2)}{(1-r^2)^2C_1^4}$$
$$= \frac{C_1^2[1-2r+r^2]}{(1-r^2)^2/C_1^4}$$
$$= \frac{1}{(1-r^2)C_1^2} \tag{15.44}$$

Substituting Eq. 15.44 into Eq. 15.42 yields

$$\mathrm{Var}(b_1) = \frac{s_e^2}{(1-r^2)\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2}. \tag{15.45}$$

Similarly, it can be proved that

$$\mathrm{Var}(b_2) = \frac{s_e^2}{(1-r^2)\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}. \tag{15.46}$$

Equations 15.45 and 15.46 are Eqs. 15.23 and 15.24, respectively.

If the correlation coefficient between x_1 and x_2—that is, r is equal to zero, then $\mathrm{Var}(b_1)$ and $\mathrm{Var}(b_2)$ reduce to

$$\mathrm{Var}(b_1) = \frac{s_e^2}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2} \quad \text{and} \quad \mathrm{Var}(b_2) = \frac{s_e^2}{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}$$

This implies that the sample variance of multiple regression slopes reduces to a simple regression case, as indicated in Eqs. 14.1 and 14.2.

Appendix 2: Derivation of Equation 15.30

From Eq. 15.27, we have

$$\hat{y}_{n+1} = a + b_1x_{1,n+1} + b_2x_{2,n+1}$$
$$= \bar{y} + b_1(x_{1,n+1}-\bar{x}_1) + b_2(x_{2,n+1}-\bar{x}_2) \tag{15.47}$$

Hence, we obtain

$$\begin{aligned}\mathrm{Var}(\hat{y}_{n+1}) &= \mathrm{Var}[\bar{y} + b_1(x_{1,n+1} - \bar{x}_1) + b_2(x_{2,n+1} - \bar{x}_2)] \\ &= \mathrm{Var}(\bar{y}) + \mathrm{Var}[b_1(x_{1,n+1} - \bar{x}_1) + \mathrm{Var}[b_2(x_{2,n+1} - \bar{x}_2)] \\ &\quad + 2\mathrm{Cov}[b_1(x_{1,n+1} - \bar{x}_1), b_2(x_{1,n+1} - \bar{x}_2)] \\ &= \frac{\sigma_{\in}^2}{n} + (x_{1,n+1} - \bar{x}_1)^2\mathrm{Var}(b_1) + (x_{2,n+1} - \bar{x}_2)^2\mathrm{Var}(b_2) \\ &\quad + 2(x_{1,n+1} - \bar{x}_1)(x_{2,n+1} - \bar{x}_2)\mathrm{Cov}(b_1, b_2) \end{aligned} \tag{15.48}$$

From Eqs. 15.40 and 15.41 in Appendix 1, we can show that ϵ

$$\begin{aligned}\mathrm{Cov}(b_1, b_2) &= E(b_1 - \beta_1)(b_2 - \beta_2) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} B_{1i}B_{2j}E(e_i e_j) \end{aligned} \tag{15.49}$$

Because $E(e_i e_j) = 0$ when $i \neq j$, Eq. 15.49 reduces to

$$\mathrm{Cov}(b_1, b_2) = s_e^2 \sum_{j=1}^{n} B_{1i}B_{2i} \tag{15.50}$$

where B_{1i} and B_{2i} are defined in Eqs. 15.38 and 15.39. It can be shown that

$$\sum_{i=1}^{n} B_{1i}B_{2i} = \frac{-r}{(1 - r^2)C_1C_2} \tag{15.51}$$

and substituting Eq. 15.51 into Eq. 15.50 yields

$$\mathrm{Cov}(b_1, b_2) = \frac{-rs_e^2}{(1 - r^2)\sqrt{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2,}i - \bar{x}_2)^2}} \tag{15.52}$$

Substituting Eqs. 15.40, 15.44, and 15.51 into Eq. 15.48, we have

$$\begin{aligned}\mathrm{Var}(y_{n+1}) = \Bigg(&\frac{1}{n} + \frac{(x_{1,n+1} - \bar{x}_1)^2}{(1 - r^2)\sum_{i=1}^{n}(x_{1,i} - \bar{x}_2)^2} + \frac{(x_{2,n+1} - \bar{x}_2)^2}{(1 - r^2)\sum_{i=1}^{n}(x_{2,i} - \bar{x}_2)^2} \\ &- \frac{2(x_{1,n+1} - \bar{x}_1)(x_{2,n+1} - \bar{x}_2)r}{(1 - r^2)\sqrt{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2,i} - \bar{x}_2)^2}}\Bigg) \end{aligned} \tag{15.53}$$

This is Eq. 15.30.

Chapter 16
Other Topics in Applied Regression Analysis

Chapter Outline

Key Terms

Multicollinearity	Proxies
Coefficient of determination	Law of diminishing returns
Heteroscedasticity	Log-log linear model
Variance inflationary factor	Log-linear model
Autocorrelation	The Durbin *H*
Durbin–Watson statistic	Dummy variable
First-order autocorrelation	Interaction
Specification error	Dynamic ratio analysis
	Term structure of interest rate

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_16, © Springer Science+Business Media New York 2013

16.1 Introduction

In Chaps. 13, 14, and 15, we discussed in some detail the technique of regression analysis and its applications. The main objectives in fitting a regression equation are (1) to estimate the regression coefficients and related parameters and (2) to predict the value of the dependent variable in terms of that of the independent variable (or variables). Several alternative specifications are possible in this kind of applied regression analysis, and a number of problems may occur.

In this chapter, we examine some of the problems associated with applying the multiple regression model. We also explore such related topics as lagged dependent variables and nonlinear regressions. Problems with the error term—specifically, violations of the assumptions of the regression model that were cited in Chaps. 14 and 15—can arise when we are running a regression. In this chapter, we discuss the detection of these problems, which include errors that are correlated and errors whose means and variance are not constant. Another problem we may encounter is a high correlation between independent variables. This problem can increase the value of standard errors and reduce the t statistics of the parameters, leading to incorrect inferences in hypothesis testing.

Other topics we address in this chapter include specification bias and model building. We also show how a nonlinear functional form can be transformed into a linear regression analysis. In some cases, for example, both independent and dependent variables can be transformed by using logarithms, and the nonlinear relationship then becomes a linear relationship. Furthermore, a regression can have a lagged dependent variable as one of the independent variables when there is a relationship between previous observations in a time series and the value in the present period. In addition, regression with dummy and interaction variables is discussed in detail. Finally, the effect of alternative business strategies is investigated.

16.2 Multicollinearity

16.2.1 Definition and Effect

The term *multicollinearity* refers to the effect, on the precision of regression parameter estimates, of two or more of the independent variables being highly correlated. For example, multicollinearity would be a problem if we were studying the cross-sectional relationship by regressing price per share against dividend per share and retained earnings per share, as discussed in Application 15.3 in Chap. 15. Because dividend per share and retained earnings per share are highly correlated, the precision of the least-squares estimated regression coefficient might be affected.

If a set of independent variables is perfectly correlated, the least-squares approach cannot be used to estimate the regression coefficients: the normal equations are not solvable. If independent variables move together, it is impossible to distinguish the separate effects of these variables on y. Perfect multicollinearity would occur,

for example, if the following independent variables were specified to model the expenditures on food for a cross section of individuals.

$$x_1 = \text{average income in dollars}$$
$$x_2 = \text{average income in cents}$$

The variables x_1 and x_2 are perfectly correlated because $x_2 = 100$ times x_1 for each of the individuals in the data set. If both of these variables were included in a regression model, least-squares results would not be obtainable because the two variables measure the same thing. Remember that the regression coefficient of x_2 is a slope term that measures the change in the dependent variable that is associated with a 1-unit change in x_2, other variables being held constant. But here it is impossible to keep the rest of the variables constant, because x_1 changes in the same direction and with the same magnitude as x_2. The solution? Simply delete one of the variables and run the regression again.

Unfortunately, most of the problems researchers face are not so easy to detect. Observations are more likely to be *highly* correlated than perfectly correlated. In such cases, least-squares estimates can be obtained but are difficult to interpret.

For example, suppose national income in period t, y_t is modeled with independent variables x_{1t} = output of manufactured goods in period t and x_{2t} = output of durable goods in period t as defined in Eq. 16.1:

$$y_t = a + b_1 x_{1t} + b_2 x_{2t} + e_t \tag{16.1}$$

If the simple correlation r between x_{1t} and x_{2t} is .90, it can be concluded that the explanatory values of the two variables overlap considerably, probably because they are highly correlated and tend to measure the same thing.

In Eq. 16.1, the first coefficient of two highly correlated variables is the slope term b_1, which measures the change in the national income that is due to a 1-unit change in the output of manufactured goods, the output of durable goods being held constant. When one of the correlated variables changes, the other is likely to change in the same direction and with approximately the same magnitude. However, the standard error of the coefficient will tend to be great, leading to lower t-values for the coefficient.[1] This increase in the standard error results from the fact that estimates are sensitive to any changes in observations or model specification.

From Eqs. 15.23 and 15.24, the sample variances of b_1 and b_2 ($S^2_{b_1}$ and $S^2_{b_2}$) of Eq. 16.1 can be defined as

[1] If x_1 and x_2 are highly correlated, then the regression can give weight to either x_1 or x_2, and it won't matter. Sampling idiosyncracies determine the choice. Hence, the large sampling error occurs.

$$S_{b_1}^2 = \frac{S_e^2}{(1 - r^2)\sum_{t=1}^{n}(x_{1t} - \bar{x}_1)^2} \tag{16.2}$$

$$S_{b_2}^2 = \frac{S_e^2}{(1 - r^2)\sum_{t=1}^{n}(x_{2t} - \bar{x}_2)^2} \tag{16.3}$$

where S_e^2 is equal to sample variance of e_t; x_1 and x_2 are means of x_{1t} and x_{2t}, respectively; and r is the correlation coefficient between x_{1t} and x_{2t}. In Eqs. 16.2 and 16.3, the factor $(1 - r^2)$ can be used to measure the impact of collinearity on $S_{b_1}^2$ and $S_{b_2}^2$. If x_{1t} and x_{2t} are uncorrelated, then $(1 - r^2) = 1$. If x_{1t} and x_{2t} are perfectly correlated ($r^2 = 1$), then $(1 - r^2) = 0$. In this case, the denominators of both $S_{b_1}^2$ and $S_{b_2}^2$ vanish, and both $S_{b_1}^2$ and $S_{b_2}^2$ equal infinity. In our national-income example, $r^2 = .81$ and $(1 - r^2) = .19$. Therefore, the precision of estimated $S_{b_1}^2$ and $S_{b_2}^2$ is greatly affected by the collinearity between x_{1t} and x_{2t}.

16.2.2 Rules of Thumb in Determining the Degree of Collinearity

Several rules of thumb are helpful when we are testing for multicollinearity. These rules involve inspection of the correlation between the independent variables. First, multicollinearity is a problem if the correlation coefficient between any two independent variables is greater than .80 or .90. If there are more than two independent variables in a regression, as indicated in Eq. 16.4, then the simple correlation coefficient between any two independent variables is not sufficient to detect the existence of multicollinearity of a regression:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \in_i \tag{16.4}$$

We must also consider that simple correlation coefficients generally fail to take into account the possible correlation between any one independent variable and all others taken as a group. Therefore, it is customary to regress each of the independent variables against all others and to note whether any of the resulting R^2-values are near 1. Using Eq. 16.4 as an example, we can define three multiple regressions in terms of x_{1i}, x_{2i}, and x_{3i}:

$$x_{1i} = a_0 + a_1 x_{2i} + a_2 x_{3i} \tag{16.5a}$$

$$x_{2i} = b_0 + b_1 x_{1i} + b_2 x_{3i} \tag{16.5b}$$

$$x_{3i} = c_0 + c_1 x_{1i} + c_2 x_{2i} \tag{16.5c}$$

$R_i^2 (i = 1, 2, 3)$ of these three regressions represents the *coefficient of determination* for the ith independent variable. It can be used to determine whether multicollinearity plagues Eq. 16.4.

In sum, to check for multicollinearity, we first calculate the three simple correlation coefficients between x_1 and x_2 (r_{12}), between x_1 and x_3 (r_{13}), and between x_2 and x_3 (r_{23}). Then we find R^2 associated with Eqs. 16.5a, 16.5b, and 16.5c. In a sense, these two methods are similar, but the first is easier to understand. This estimated information can be used to determine the existence of multicollinearity.

One way to measure collinearity is to use the measurements of $(1 - R^2)$ indicated in Eq. 16.2 or 16.3 to construct a *variance inflationary factor* (VIF) for each explanatory variable in Eq. 16.4:

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \tag{16.6}$$

where R_i^2 is as defined in Eqs. 16.5a, b, and c. If there are only two independent variables, R_i is merely the correlation coefficient. If a set of independent variables is uncorrelated, then VIF_i is equal to 1. If R_i^2 approaches 1, both VIF and the standard deviations of the slopes ($S_{b_1}^2$ and $S_{b_2}^2$) approach infinity. Researchers have used $\text{VIF}_i = 10$ as a critical-value rule of thumb to determine whether too much correlation exists between the ith independent variable and other independent variables.[2] The corresponding R_i values of VIF_i at least 10 is now illustrated. Thus, $1/(1 - R_i^2) \geq 10$ implies $1 - R_i^2 \leq .1$. This implies that $.95 \leq R_i \leq 1$ or $-1 \leq R_i \leq -.95$.

Example 16.1 Analyzing the Determination of Price per Share. Data on dividend per share (DPS), retained earnings (RE) per share, and price per share (PPS) from 2007 to 2009 for the 30 firms employed to compile the Dow Jones Industrial Average are used to estimate Eq. 16.7. Regression results are shown in Table 16.1

$$\text{PPS}_{i,t} = a + b(\text{RE}_{i,t}) + c(\text{DPS}_{i,t}) + e_{i,t} \tag{16.7}$$

where $\text{PPS}_{i,t}$, $\text{DPS}_{i,t}$, and $\text{RE}_{i,t}$ represent price per share, dividends per share, and retained earnings per share for the ith firm in the tth year, respectively. This cross-sectional model states that the price per share is a function of dividends per share and retained earnings per share. The results of the regressions seem to be satisfactory with all of the independent variables and appear significant at the 5 % level ($F_{.05,2,27} = 3.35$, from Table A6). However, we must examine the correlations between $\text{DPS}_{i,t}$ and $\text{RE}_{i,t}$ to determine whether multicollinearity may be a problem.

The 2009 regression results indicated in column (4) of Table 16.1 are used to determine the degree of multicollinearity. To do this analysis, we assemble the

[2] See Marquardt, D.W.: You should standardize the prediction variables in your regression models, discussion of A Critique of Some Ridge Regression Methods, by G. Smith and F. Campbell, J. Am. Stat. Assoc. **75**, 87–91 (l980).

Table 16.1 Regression results of Eq. 16.7

(1)	(2) 2007	(3) 2008	(4) 2009
Constant	19.413	14.536	10.465
	(3.46)	(3.26)	(2.23)
RE	8.14	6.389	7.605
	(6.20)	(7.79)	(5.57)
DPS	5.47	2.208	10.581
	(1.15)	(0.61)	(2.45)
R^2	.70	.77	.77
F-statistic	31.33	44.46	45.53

The t-values are indicated in parentheses

Table 16.2 A simple correlation matrix

Variable	PPS, y	RE, x_1	DPS, x_2
PPS, y	$r_{y,y} = 1.000$	$r_{y,1} = .8488$	$r_{y,2} = .7132$
RE, x_1	—	$r_{1,1} = 1.000$	$r_{1,2} = .6351$
DPS, x_2	—	—	$r_{2,2} = 1.000$

correlation coefficients among PPS, DPS, and RE in Table 16.2. We note from the correlation matrix that each variable is perfectly correlated with itself; hence, we find three entries equal to 1.000 along the diagonal of the table. It is the red-colored entry that is of crucial importance. Substituting $r_{1,2} = .6351$ into Eq. 16.6, we obtain $VIF_1 = 1/[1 - (.6351)^2] = 1.67603$. Because 1.67603 is much smaller than 10, we conclude that the degree of collinearity for this regression is relatively unimportant.

16.3 Heteroscedasticity

16.3.1 Definition and Concept

Heteroscedasticity arises when the variances of the error terms of a regression model are not constant over different sample observations. For example, heteroscedasticity would be a problem in a study of sales for a cross section of firms in an industry, because the error terms for large firms would be likely to have larger variances than those for small firms. In other words, the high volatility in sales for larger firms might pose problems for the researcher. The probable error terms are shown in Fig. 16.1. This figure indicates that the magnitude of error terms is a function of firm size. For example, it may be the case that $\sigma_i^2 < \sigma_j^2 < \sigma_k^2$.

Another commonly cited example of heteroscedasticity is the relationship between family expenditures and income. High-income families are likely to exhibit a higher variance in spending than lower-income families. Figure 16.2 shows a reasonable plot of level of income versus level of expenditures. This figure indicates that expenditures for consumers with lower income have smaller error terms.

Heteroscedasticity poses a problem when we are estimating parameters in the regression model, because the least-squares estimation procedure places more weight

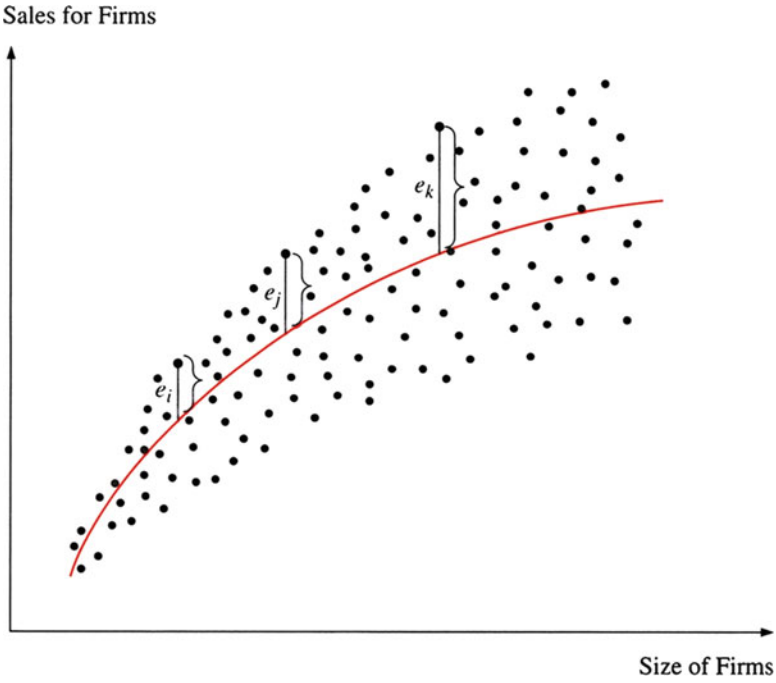


Fig. 16.1 A possible relationship between sales and firm size

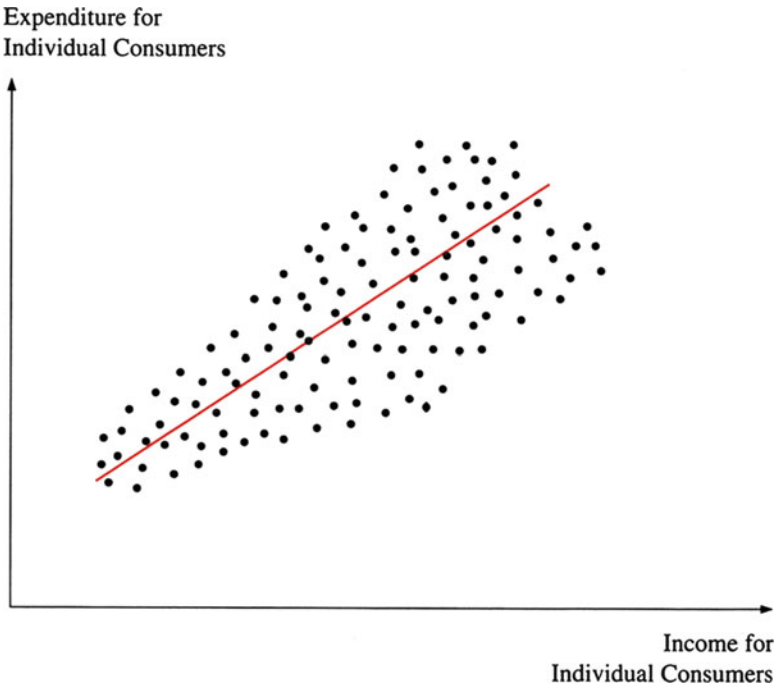


Fig. 16.2 Cross-sectional relationship between expenditure and income

on observations that have large errors and variances. Thus, the regression line is adjusted to give a good fit for the large-variance portion of the observations but largely ignores the small-variance part of the data. The result is that the variances of the estimates do not have a minimum variance.

16.3.2 Evaluating the Existence of Heteroscedasticity

The easiest way to check for heteroscedasticity is to look at a plot of the residuals against the independent variables or the expected values. To estimate the error term, we first compute the predicted dependent value by the regression model

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} \tag{16.8}$$

Then, we calculate the error term by taking the actual value for y_i and subtracting the predicted value: $e_i = y_i - \hat{y_i}$. In practice, we generally plot the residuals range e_i against independent variables on a series of graphs or examine the predicted values $\hat{y_i}$. If the residuals appear to be random and the width of the scatter diagram seems constant throughout the data—that is, if no pattern is apparent—then no heteroscedasticity is present.

A somewhat more involved evaluation for the existence of heteroscedasticity consists of the following steps:

1. Run a standard regression.
2. Calculate the residuals, $e_i = \hat{y_i} - y_i$.
3. Run a regression using the square of the residuals as the dependent variable and the estimated dependent variable $\hat{y_i}$ as the independent variable.
4. Estimate nR^2, where n is the sample size and R^2 is the coefficient of determination.
5. Use the χ^2 statistic with 1 degree of freedom to test whether nR^2 is significantly different from zero.

The use of this method to analyze the existence of heteroscedasticity is illustrated in the following example.

Example 16.2 Residual Heteroscedasticity Analysis for Price per Share. The regression results for Eq. 16.7 were run for the 30 Dow Jones industrials for 2007, 2008, and 2009. These results appear in Table 16.1.

To check for heteroscedasticity, we calculate the residuals of the regression and plot them against the predicted values $\hat{y}_{i,t}$. Residuals (e) and predicted PPS values ($\hat{y}$) for all three years are listed in Table 16.3. Figure 16.3 shows the plots of residuals from the least-squares regression against $\hat{y}_{i,t}$ for 2007. Similar plots for 2008 and 2009 are given in Figs. 16.4 and 16.5. As can be seen in all three plots, there are patterns in the residuals. The pattern for 2007 is stronger than those for 2008 and 2009. We might conclude that the residuals plotted against predicted values are not random and do violate the standard assumptions of the regression model.

Table 16.3 Residuals (e) and predicted PPS values (ŷ) for 30 Dow Jones Industrials

Year	2007		2008		2009	
Observations	ŷ	e	ŷ	e	ŷ	e
1	47.38	−10.83	47.38	−10.83	4.84	11.28
2	50.77	1.25	50.77	1.25	29.87	10.65
3	43.87	−0.18	43.87	−0.18	39.90	−6.77
4	70.69	16.77	70.69	16.77	42.62	11.51
5	71.79	0.77	71.79	0.77	39.27	17.72
6	63.99	−20.34	63.99	−20.34	33.19	8.48
7	103.62	−10.29	103.62	−10.29	78.61	−1.62
8	47.93	13.44	47.93	13.44	50.25	6.75
9	40.07	−5.68	40.07	−5.68	27.71	−0.25
10	54.17	−10.08	54.17	−10.08	42.50	−8.83
11	86.80	6.89	86.80	6.89	58.37	9.82
12	43.52	−6.45	43.52	−6.45	26.98	−11.85
13	43.62	8.06	43.62	8.06	38.26	9.20
14	45.89	−5.15	45.89	−5.15	22.40	−2.00
15	31.64	−4.98	31.64	−4.98	110.18	20.72
16	87.18	20.92	87.18	20.92	64.73	−0.32
17	58.14	8.56	58.14	8.56	63.87	−1.43
18	43.57	15.34	43.57	15.34	69.67	−33.13
19	37.93	20.18	37.93	20.18	66.73	15.94
20	76.30	8.02	76.30	8.02	8.68	6.38
21	59.80	−18.54	59.80	−18.54	28.28	−10.09
22	35.44	−12.71	35.44	−12.71	56.41	−5.31
23	52.62	8.57	52.62	8.57	43.94	−15.91
24	43.05	−1.49	43.05	−1.49	58.47	10.94
25	61.45	15.09	61.45	15.09	28.15	−4.38
26	45.92	1.77	45.92	1.77	18.45	3.56
27	33.26	−3.79	33.26	−3.79	72.00	−22.14
28	29.26	−0.35	29.26	−0.35	38.25	−11.07
29	82.88	−29.08	82.88	−29.08	46.07	1.05
30	38.34	−5.71	38.34	−5.71	30.41	−8.88

In addition to making this visual inspection of residuals, we must run a regression with the squared error terms as the dependent variable and $\hat{y_i}$ as the independent variable. Let's look at the results for 2009:

$$e_{i,t}^2 = c + f\hat{y}_{i,t} + \text{residuals} \tag{16.9}$$

Using $e^2{}_{i,t}$ and $\hat{y_{i,t}}$ for 2009 as presented in Table 16.4, we estimate Eq. 16.9 and obtain

$$e_{i,2009}^2 = -39.1519 + 4.0942\hat{y}_{i,2009} \qquad R^2 = .1758$$
$$(t = 2.44)$$

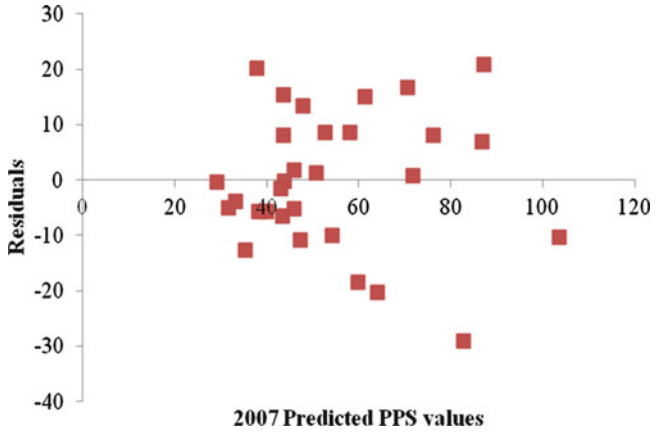


Fig. 16.3 Plots of residuals against the predicted PPS values, $\hat{y}$ (2007)

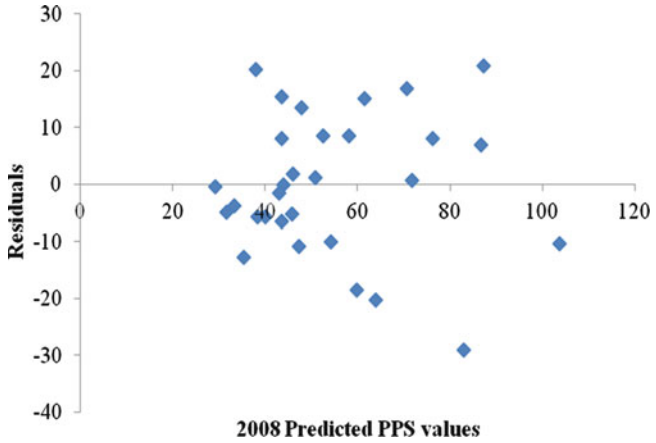


Fig. 16.4 Plots of residuals against the predicted PPS values, $\hat{y}$ (2008)

Because there are $n = 30$ sets of observations, the test is based on $nR^2 = (30)(.1758) = 5.2740$. From Table A5 in Appendix A, we find that for a test at the 5 % level, $\chi^2_{1,0.05}$=3.84. Therefore, we can conclude that the residuals in the regression of price per share on dividends per share and retained earnings per share do not have the same variance, and the null hypothesis should be rejected. One way to deal with this problem is to use a two-stage procedure to estimate the parameters of regression models. In the first stage, we estimate the parameters of Eq. 16.7 and the predicted value $\hat{y}_{i,t}$ of the dependent variable. Predicted values of $y_{i,t}$ for 2009 are listed in Table 16.4. In the second stage, we estimate a transformed Eq. 16.7:

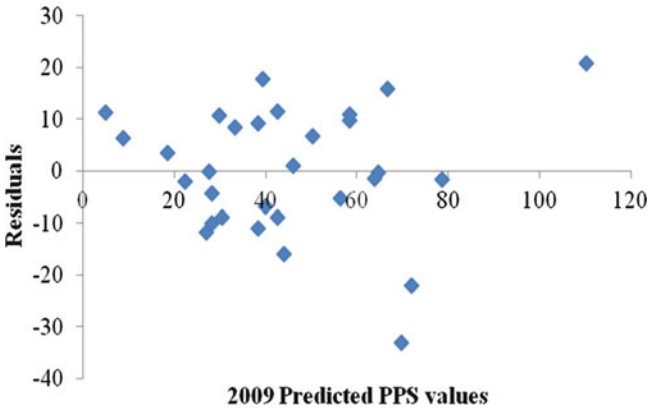


Fig. 16.5 Plots of residuals against the predicted PPS values, $\hat{y}$ (2009)

Table 16.4 $e^2_{i,t}$ and $\hat{y}_{i,t}$ for a test of heteroscedasticity for Eq. 16.7 in terms of 2009 data

$\hat{y}_{i,2009}$	$e_{i,2009}$	$e^2_{i,2009}$
4.84	11.28	127.3021
29.87	10.65	113.3852
39.90	−6.77	45.88082
42.62	11.51	132.5925
39.27	17.72	314.0423
33.19	8.48	71.99254
78.61	−1.62	2.638521
50.25	6.75	45.51427
27.71	−0.25	0.060564
42.50	−8.83	77.9001
58.37	9.82	96.33882
26.98	−11.85	140.305
38.26	9.20	84.5601
22.40	−2.00	3.99501
110.18	20.72	429.3093
64.73	−0.32	0.102427
63.87	−1.43	2.045569
69.67	−33.13	1097.628
66.73	15.94	254.0671
8.68	6.38	40.66668
28.28	−10.09	101.8989
56.41	−5.31	28.23771
43.94	−15.91	253.1638
58.47	10.94	119.5953
28.15	−4.38	19.20552
18.45	3.56	12.66711
72.00	−22.14	490.2615
38.25	−11.07	122.6317
46.07	1.05	1.100324
30.41	−8.88	78.80692

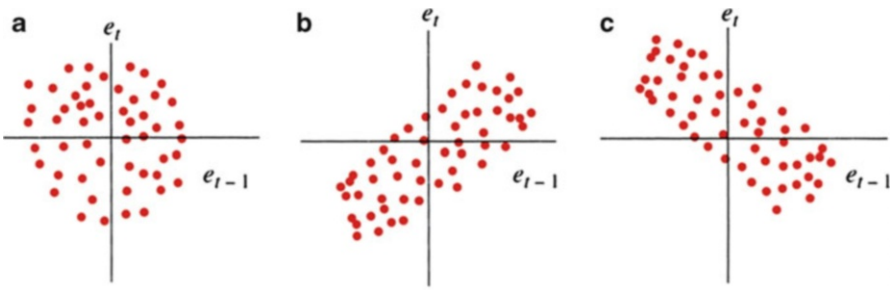


Fig. 16.6 (**a**) No autocorrelation (**b**) Positive autocorrelation (**c**) Negative autocorrelation

$$\frac{y_{i,t}}{\hat{y}_{i,t}} = a\frac{1}{\hat{y}_{i,t}} + b\frac{x_{1i,t}}{\hat{y}_{i,t}} + c\frac{x_{2i,t}}{\hat{y}_{i,t}} + e'_{i,t} \qquad (16.7')$$

where $e'_{i,t}$ is an error term that approximates constant variance. Using 2009 data, we estimate Eq. 16.7′ and its results:

$$\frac{\hat{y}_{i,2009}}{\hat{y}_{i,2009}} = \underset{(9.87)}{17.333}\left(\frac{1}{\hat{y}_{i,2009}}\right) + \underset{(4.15)}{4.553}\frac{x_{1i,2009}}{\hat{y}_{i,2009}} + \underset{(2.92)}{10.316}\frac{x_{2i,2009}}{\hat{y}_{i,2009}}$$

From the *t*-values indicated in parentheses, we find that the *t*-values for second-stage estimates are more efficient than those for the one-stage estimates indicated in Table 16.1.

16.4 Autocorrelation

16.4.1 Basic Concept

One of the assumptions of the regression model is that the errors are uncorrelated; in other words, the correlation between error terms is equal to zero. We are quite likely to encounter only uncorrelated errors when dealing with cross-sectional data. However, autocorrelation—the correlation of an error term and a lagged version of itself—is likely to occur with time-series data because errors made in a particular time period are readily carried over to future time periods. For example, an underestimate of the GDP in one year can generate more underestimates in future time periods.

Figure 16.6 shows examples of autocorrelation. The error in period t is graphed on the y axis, the error in period $t - 1$ on the x axis. If no autocorrelation exists, the plot has no trend, as shown in Fig. 16.6a. The upward slope of the diagram in Figure 16.6b signals positive autocorrelation. This implies that a large error in period $t - 1$ will be associated with a large error in period t. Negative

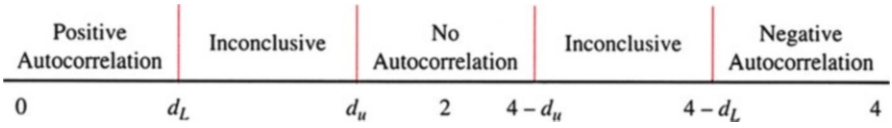


Fig. 16.7 Critical values of the Durbin–Watson statistic

autocorrelation is apparent in Fig. 16.6c, where high errors in the previous period tend to result in low errors in the next period. Again, if plotting reveals a pattern, the regression assumption that the errors are random and not correlated through time has been violated.

16.4.2 The Durbin–Watson Statistic

Detecting autocorrelation by inspecting errors is difficult; thus, the *Durbin–Watson statistic* (DW) is generally used to detect first-order autocorrelation. *First-order autocorrelation* occurs when correlation between errors is separated by one period. Here, we will discuss both first-order autocorrelation and the Durbin–Watson (DW) statistic in detail.

If e_t and e_{t-1} are residual terms in periods t and $t-1$, respectively, then first-order correlation r_1 for these error terms is defined as follows:

$$r_1 = \frac{\sum_{t=2}^{n} (e_t - \bar{e})(e_{t-1} - \bar{e})}{\sum_{t=1}^{n} (e_t - \bar{e})^2} \tag{16.10}$$

where $\bar{e} = \sum_{t=1}^{n} e_t / n$.

The DW statistic can be used to test the null hypothesis that no first-order autocorrelation exists among the residuals of a regression. The statistic, calculated from the residuals, is

$$\text{DW} = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{16.11}$$

where e_t and e_{t-1} are error terms in periods t and $t-1$, respectively. This statistic is calculated by summing the difference in the error terms separated by one period and dividing by the squared error term. Figure 16.7 shows that the DW statistic falls between zero and 4. If no autocorrelation exists, DW equals 2, because the difference in the error terms is directly proportional to the error term in period t.

Positive autocorrelation exists if DW is low, because the difference between the error term in period t and in period $t-1$ tends to be very small. Negatively correlated errors are closer to 4 because the difference in the error terms tends to be large.

We can use the DW statistics listed in the Table A9 of Appendix A to determine whether the null hypothesis is accepted or rejected. These tables give two values, d_L and d_U, for different sample sizes n and number of independent variables k. If the DW statistic falls between d_U and $4-d_U$, the null hypothesis that the correlation between lagged errors is equal to zero is accepted. If the statistic is less than d_L, that null hypothesis is rejected in favor of positive autocorrelation. Negative autocorrelation exists if the DW is greater than $4-d_L$. Did you notice that there is an indeterminate zone in which no judgment can be made? This zone is between d_L and d_U or $4-d_U$ and $4-d_L$. (The probability level of d_U, d_L depends on whether we are performing a one- or a two-tailed test.)

Example 16.3 How to Detect First-Order Autocorrelation. Annual rates of return for both JNJ and MRK and market rates of return during 1990 to 2009, which can be found in Table 4.15 in Chap. 4, are used to estimate the market model of Eq. 16.12 for JNJ and MRK:

$$R_{i,t} = a_i + b_i R_{m,t} + e_{i,t} \tag{16.12}$$

where $R_{i,t}$ is the rate of return for the ith firm in period t, $R_{m,t}$ is the market rate of return, a_i and b_i are regression parameters, and $e_{i,t}$ is the error term.

MINITAB outputs of the estimated market models for JNJ and MRK are presented in Figs. 16.8 and 16.9. These two outputs indicate that the beta coefficients b_i for JNJ and MRK are .639 and .7886, respectively. In addition, we see that the DW statistics for JNJ and MRK are 2.5128 and 2.45845. Remember, this test determines whether there is evidence of autocorrelation among the residuals. In this example, there are one independent variable and 20 observations. Looking these values up in Table A9 of

Regression Analysis: JNJ versus S&P

The regression equation is

JNJ = 0.0273 + 0.639 S&P

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.02726	0.07227		0.38	0.710
S&P	0.6386	0.4472	1.43	0.170	1.000

S = 0.294804 R-Sq = 10.2% R-Sq(adj) = 5.2%

Analysis of Variance

Fig. 16.8 MINITAB output of market model for JNJ

Source	DF	SS	MS	F	P
Regression	1	0.17720	0.17720	2.04	0.170
Residual Error	18	1.56437	0.08691		
Total	19	1.74158			

Obs	S&P	JNJ	Fit	SE Fit	Residual	St Resid
1	0.036	0.2301	0.0505	0.0673	0.1796	0.63
2	0.124	0.6168	0.1066	0.0709	0.5102	1.78
3	0.105	-0.5513	0.0944	0.0682	-0.6457	-2.25R
4	0.086	-0.0916	0.0821	0.0665	-0.1736	-0.60
5	0.020	0.2449	0.0400	0.0691	0.2049	0.71
6	0.177	0.5846	0.1400	0.0823	0.4445	1.57
7	0.238	-0.4098	0.1791	0.1011	-0.5888	-2.13R
8	0.303	0.3408	0.2205	0.1246	0.1203	0.45
9	0.243	0.2877	0.1823	0.1029	0.1054	0.38
10	0.223	0.1242	0.1695	0.0962	-0.0453	-0.16
11	0.075	0.1397	0.0753	0.0660	0.0644	0.22
12	-0.163	-0.4312	-0.0770	0.1220	-0.3542	-1.32
13	-0.168	-0.0780	-0.0798	0.1236	0.0018	0.01
14	-0.029	-0.0212	0.0088	0.0785	-0.0300	-0.11
15	0.171	0.2486	0.1367	0.0810	0.1119	0.39
16	0.068	-0.0325	0.0705	0.0659	-0.1030	-0.36
17	0.086	0.1225	0.0819	0.0665	0.0406	0.14
18	0.127	0.0346	0.1085	0.0713	-0.0739	-0.26
19	-0.174	-0.0764	-0.0839	0.1261	0.0075	0.03
20	-0.223	0.1085	-0.1151	0.1452	0.2236	0.87

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 2.51280

Fig. 16.8 (continued)

Appendix A for a level of significance of 5 % (under a two-tailed test, $\alpha = 10\,\%$), we find that d_L is 1.20 and d_U is 1.41. Because both DW values are greater than d_U and less than $4 - d_U$, we conclude that there is no evidence of autocorrelation in either regression.

Dividend per share (DPS_t) and earnings per share (EPS_t) for both JNJ and MRK during 1990–2009 (presented in Table 2.3 in Chap. 2) are used to estimate the regression specified in Eq. 16.13:

Regression Analysis: MRK versus S&P

The regression equation is

MRK = 0.035 + 0.789 S&P

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.0348	0.1027	0.34	0.738	
S&P	0.7886	0.6353	1.24	0.230	1.000

S = 0.418787 R-Sq = 7.9% R-Sq(adj) = 2.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.2702	0.2702	1.54	0.230
Residual Error	18	3.1569	0.1754		
Total	19	3.4271			

Obs	S&P	MRK	Fit	SE Fit	Residual	St Resid
1	0.036	0.1854	0.0635	0.0955	0.1219	0.30
2	0.124	0.8786	0.1328	0.1007	0.7458	1.83
3	0.105	-0.7338	0.1178	0.0969	-0.8516	-2.09R
4	0.086	-0.1830	0.1025	0.0945	-0.2855	-0.70
5	0.020	0.1424	0.0506	0.0981	0.0919	0.23
6	0.177	0.7539	0.1741	0.1170	0.5798	1.44
7	0.238	0.2353	0.2223	0.1437	0.0130	0.03
8	0.303	0.3525	0.2735	0.1770	0.0791	0.21
9	0.243	0.4097	0.2263	0.1461	0.1834	0.47
10	0.223	-0.5371	0.2105	0.1366	-0.7476	-1.89
11	0.075	0.4119	0.0942	0.0938	0.3177	0.78

Fig. 16.9 MINITAB output market model for MRK

12	-0.163	-0.3574	-0.0939	0.1733	-0.2635	-0.69
13	-0.168	-0.0133	-0.0974	0.1756	0.0841	0.22
14	-0.029	-0.1583	0.0120	0.1114	-0.1703	-0.42
15	0.171	-0.2720	0.1700	0.1150	-0.4419	-1.10
16	0.068	0.0369	0.0882	0.0936	-0.0513	-0.13
17	0.086	0.4183	0.1023	0.0944	0.3161	0.77
18	0.127	0.3674	0.1352	0.1013	0.2323	0.57
19	-0.174	-0.4508	-0.1025	0.1791	-0.3483	-0.92
20	-0.223	0.2541	-0.1410	0.2062	0.3950	1.08

R denotes an observation with a large standardized residual.
Durbin-Watson statistic = 2.45845

Fig. 16.9 (continued)

$$\text{DPS}_{i,t} = a_i + b_i\text{EPS}_{i,t} + e_{i,t} \tag{16.13}$$

where $\text{DPS}_{i,t}$ and $\text{EPS}_{i,t}$ are dividend per share and earnings per share for the ith firm in period t.

MINITAB output for estimated Eq. 16.13 is presented in Figs. 16.10 and 16.11. From Figs. 16.10 and 16.11, we know that DPS is highly correlated with EPS for both JNJ and MRK. And the DW statistics are 1.77395 and 1.06240, respectively.

In a two-tailed test, we look up in the Durbin–Watson table critical values for a 5 % level of significance, the number of observations 20, and the number of independent variables 1. The critical values are 1.20 and 1.41. Remember that if the DW falls between the two values, the test is inconclusive. If it is less than 1.20, positive autocorrelation is a problem. The DW of JNJ is larger than 1.20, and the DW of MRK is below that value, so we conclude that positive autocorrelation exists among the residuals in the regression model of Eq. 16.13 for MRK.

When results of Eq. 16.13 for MRK, as indicated in Eq. 16.11, imply that the residuals of regression might be autocorrelated, least-squares estimates and inferences based on them can be very unreliable. Under these circumstances, a modified model of Eq. 16.13 can be used to adjust for the autocorrelation:

$$\text{DPS}_{i,t} - \hat{\rho}\text{DPS}_{i,t-1} = a_i(1-\hat{\rho}) + b_i(\text{EPS}_{i,t} - \hat{\rho}\text{EPS}_{i,t-1}) + e'_{i,t} \tag{16.13a}$$

where $e'_{i,t} = e_{i,t} - \hat{\rho}\, e_{i,t-1}$, $\hat{\rho} = r_1 = 1 - DW/2$ = estimated first-order autocorrelation. MINITAB output in terms of Eq. 16.13a for MRK is presented in Fig. 16.12. From this figure, we find that

$$\text{DPS}_{i,t} - \hat{\rho}\text{DPS}_{i,t-1} = .387 + .233(\text{EPS}_{i,t} - \hat{\rho}\text{EPS}_{i,t-1})$$
$$(t = 5.26) \qquad \text{DW} = 1.39$$

This result implies that the DW statistic has improved substantially.

16.5 Model Specification and Specification Bias (Optional)

A specification error is the error associated with either omitting a relevant variable from a regression model or including an irrelevant variable in it. When specifying a regression model (i.e., when determining which variables should be included in the model), we must make two decisions: *which* variables to include and *what* functional form—a log form, a squared term, or a lagged term—to use.

Data Display

Row	EPS(JNJ)	lagEPS(JNJ)	DPS(JNJ)	lagDPS(JNJ)
1	3.38	3.19	1.29	1.10
2	4.30	3.38	1.51	1.29
3	1.54	4.30	0.88	1.51
4	2.71	1.54	1.00	0.88
5	3.08	2.71	1.12	1.00
6	3.65	3.08	1.25	1.12
7	2.12	3.65	0.72	1.25
8	2.41	2.12	0.83	0.72
9	2.23	2.41	0.95	0.83
10	2.94	2.23	1.04	0.95
11	3.39	2.94	1.22	1.04
12	1.83	3.39	0.66	1.22
13	2.16	1.83	0.78	0.66
14	2.39	2.16	0.91	0.78
15	2.83	2.39	1.08	0.91
16	3.46	2.83	1.26	1.08
17	3.73	3.46	1.44	1.26
18	3.63	3.73	1.60	1.44
19	4.57	3.63	1.77	1.60
20	4.40	4.57	1.91	1.77

Fig. 16.10 MINITAB output of Eq. 16.13 for JNJ

Fig. 16.10 (continued)

Regression Analysis: DPS(JNJ) versus EPS(JNJ)

The regression equation is

DPS(JNJ) = 0.019 + 0.376 EPS(JNJ)

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.0186	0.1025	0.18	0.858	
EPS(JNJ)	0.37612	0.03251	11.57	0.000	1.000

S = 0.123214 R-Sq = 88.1% R-Sq(adj) = 87.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2.0320	2.0320	133.84	0.000
Residual Error	18	0.2733	0.0152		
Total	19	2.3052			

Unusual Observations

Obs	EPS(JNJ)	DPS(JNJ)	Fit	SE Fit	Residual	St Resid
3	1.54	0.8769	0.5978	0.0559	0.2791	2.54R
20	4.40	1.9099	1.6735	0.0522	0.2364	2.12R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 1.77395

There is often a theoretical basis for selecting the independent variables for a regression model. For example, economic theory states that the demand for a product is a function of price, cost of substitute goods, income, and consumer tastes. In practice, of course, it may be impossible for the researcher to obtain information on all of these items, so *proxies* for the variables are used instead. Because it may be difficult to obtain data on the price of related goods, for example, some type of price index can be used as a proxy.

Model building is more of an art than science. The researcher tries to include all the variables that affect the outcome of the dependent variable, but no specification can perfectly determine the movements and attributes of the variable in question. The best the researcher can do is search for variables that seem consistent with underlying theory, practice, and common sense. Model specification is of great importance: if significant explanatory variables are left out, the model's worth is

compromised even though least-squares estimates of the parameters are obtained. Here again, good judgment and reliance on theory must guide the researcher.

Example 16.4 Impact of the Omission of Variables on Estimated Regression Coefficients. Suppose we omit retained earnings (RE) from Eq. 16.7 in Example 16.1 for the year 2009. The equation becomes

Data Display

Row	EPS(MRK)	lagEPS(MRK)	DPS(MRK)	lagDPS(MRK)
1	4.51	3.74	2.00	1.70
2	5.39	4.51	2.34	2.00
3	1.70	5.39	0.95	2.34
4	1.86	1.70	1.06	0.95
5	2.35	1.86	1.15	1.06
6	2.63	2.35	1.24	1.15
7	3.12	2.63	1.44	1.24
8	3.74	3.12	1.70	1.44
9	4.30	3.74	1.93	1.70
10	2.45	4.30	1.09	1.93
11	2.90	2.45	1.23	1.09
12	3.14	2.90	1.36	1.23
13	3.14	3.14	1.41	1.36
14	3.03	3.14	1.45	1.41
15	2.61	3.03	1.50	1.45
16	2.10	2.61	1.52	1.50
17	2.03	2.10	1.52	1.52
18	1.49	2.03	1.51	1.52
19	3.64	1.49	1.52	1.51
20	5.68	3.64	1.58	1.52

Regression Analysis: DPS(MRK) versus EPS(MRK)

The regression equation is

DPS(MRK) = 0.821 + 0.211 EPS(MRK)

Fig. 16.11 MINITAB output of Eq. 16.13 for MRK

Predictor Coef SE Coef T P VIF

Constant 0.8209 0.1509 5.44 0.000

EPS(MRK) 0.21133 0.04585 4.61 0.000 1.000

S = 0.232678 R-Sq = 54.1% R-Sq(adj) = 51.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1.1503	1.1503	21.25	0.000
Residual Error	18	0.9745	0.0541		
Total	19	2.1248			

Unusual Observations

Obs	EPS(MRK)	DPS(MRK)	Fit	SE Fit	Residual	St Resid
20	5.68	1.5836	2.0213	0.1296	-0.4377	-2.27RX

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.

Durbin-Watson statistic = 1.06240

Fig. 16.11 (continued)

$$\text{PPS}_i = a' + b'\text{DPS}_i + e'_i \tag{16.7'}$$

where

$$b' = \frac{\text{Cov}(\text{PPS}, \text{DPS})}{\text{Var}(\text{DPS})} = \frac{12.3582}{0.47814} = 25.8462$$

$$a' = \overline{\text{PPS}} - \hat{b}'\overline{\text{DPS}} = 44.6360 - (25.8462)(1.2145) = 13.2458$$

By regressing RE_i on DPS_i, we obtain the auxiliary regression:

$$\text{RE}_i = b_0 + b_1\text{DPS}_i \tag{16.14}$$

Data Display

Row	EPS(MRK)	DPS(MRK)	LolagEPS(MRK)	LolagDPS(MRK)	dif(EPS_MRK))
1	4.51	2.00	1.75130	0.79812	2.75870
2	5.39	2.34	2.11560	0.93606	3.27440
3	1.70	0.95	2.52771	1.09641	-0.82771
4	1.86	1.06	0.79565	0.44386	1.06435
5	2.35	1.15	0.87109	0.49824	1.47891
6	2.63	1.24	1.10092	0.53746	1.52908
7	3.12	1.44	1.23162	0.58272	1.88838
8	3.74	1.70	1.46125	0.67519	2.27875
9	4.30	1.93	1.75192	0.79537	2.54808
10	2.45	1.09	2.01569	0.90372	0.43431
11	2.90	1.23	1.14841	0.51261	1.75159
12	3.14	1.36	1.35901	0.57887	1.78099
13	3.14	1.41	1.46997	0.63712	1.67003
14	3.03	1.45	1.47197	0.65970	1.55803
15	2.61	1.50	1.42130	0.67928	1.18870
16	2.10	1.52	1.22409	0.70099	0.87591
17	2.03	1.52	0.98671	0.71132	1.04329
18	1.49	1.51	0.95011	0.71116	0.53989
19	3.64	1.52	0.70022	0.70776	2.93978
20	5.68	1.58	1.70632	0.71029	3.97368

Row	dif(DPS_MRK)
1	1.19859
2	1.40270
3	-0.14961
4	0.61893
5	0.64821
6	0.70555
7	0.85753
8	1.02142
9	1.13237
10	0.18972
11	0.72218

Fig. 16.12 MINITAB output of Eq. 16.13a for MRK

12 0.78017

13 0.77008

14 0.78928

15 0.81600

16 0.81633

17 0.80566

18 0.79858

19 0.80736

20 0.87329

Regression Analysis: dif(DPS_MRK) versus dif(EPS_MRK))

The regression equation is
dif(DPS_MRK) = 0.387 + 0.233 dif(EPS_MRK))

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.38673	0.08842	4.37	0.000	
dif(EPS_MRK))	0.23318	0.04433	5.26	0.000	1.000

S = 0.210802 R-Sq= 60.6% R-Sq(adj) = 58.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1.2296	1.2296	27.67	0.000
Residual Error	18	0.7999	0.0444		
Total	19	2.0295			

Unusual Observations

Obs	dif(EPS_MRK))	dif(DPS_MRK)	Fit	SE Fit	Residual	St Resid
3	-0.83	-0.1496	0.1937	0.1210	-0.3433	-1.99 X
20	3.97	0.8733	1.3133	0.1118	-0.4400	-2.46R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

Durbin-Watson statistic = 1.39124

Fig. 16.12 (continued)

where

$$b_1 = \frac{\text{Cov}(\text{RE}, \text{DPS})}{\text{Var}(\text{DPS})} = \frac{0.9597}{0.47814} = 2.0072$$

$$b = \overline{\text{RE}} - b\,\overline{\text{DPS}} = 2.8033 - (2.0072)(1.2145) = 0.3656$$

From the specification analysis of Theil (1971),[3] the relationship among b, b', c, and b_1 can be defined as

$$b' = c + b_1 b \tag{16.15}$$

where b' and b_1 are estimated in accordance with Eqs. 16.7′ and 16.14, respectively, and b and c are estimated by using Eq. 16.7. Substituting into the foregoing equation the estimated b and c (from Table 16.1), $b_1 = 2.0072$, and $b' = 25.8462$, we obtain

$$25.8462 \cong 10.581 + (2.0072)(7.605) = 25.8454$$

This implies that the misspecification error when RE_i is omitted from Eq. 16.7 is 15.2644 (25.8454–10.581) for b.

16.6 Nonlinear Models (Optional)

Thus far, we have assumed that there is a linear relationship between the dependent variable and a set of independent variables. This assumption yields a convenient approximation of the phenomena being modeled. However, there are times when other functional forms of the independent variable provide a better depiction of reality. In this section, we discuss nonlinear models, including quadratic and log-linear models. We will continue to use the same regression concepts, such as hypothesis testing and confidence intervals, in our analysis.

16.6.1 The Quadratic Model

A quadratic model takes the form

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + e_i \tag{16.16}$$

[3] Theil, H.: Principles of Econometrics, Wiley, New York (1971).

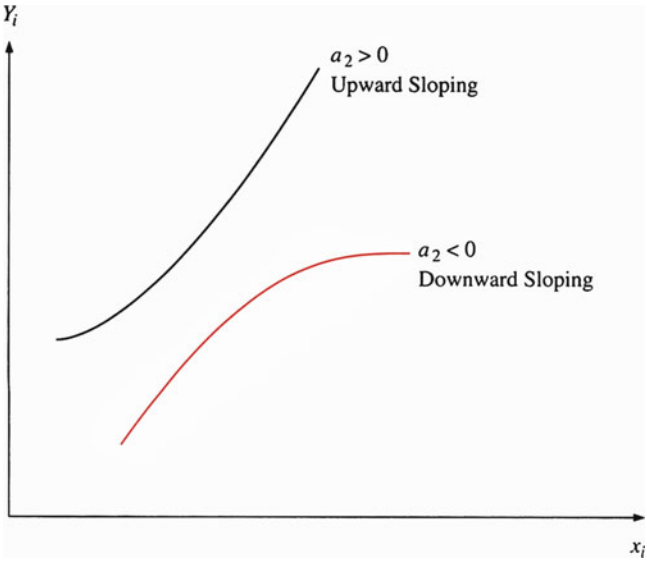


Fig. 16.13 Two different types of nonlinear curve

The only difference between this model and the models previously specified is the squared term in this model. The square of the variable is calculated and used as another independent variable, and a regression on the data is run. The quadratic term traces a parabola, as shown in Fig. 16.13. If the parameter for the squared term has a positive value, the parabola opens upward. A negative parameter implies a downward-opening parabola. The linear model uses only the part of the data that is available to fit the curve. An example is shown in Fig. 16.14. Here, only the upward-sloping part is used by the linear model to fit the data.

Again, we must exercise judgment and common sense to determine whether a quadratic term is needed in the model. We may have some idea how the dependent variable will react to changes in the independent variable, and graphs of the data may give us more information to help us specify the model. A quadratic model might be of interest in a production function where output of a product is the dependent variable and an input is the independent variable. The *law of diminishing returns* states that after a certain point, the marginal product of the variable inputs declines when additional units of a variable input are added to fixed inputs. In agriculture, for example, doubling the fertilizer doubles the output of corn at low levels of fertilizer use. However, further increases in fertilizer increase the output only marginally. If a regression were to be run, the sign of the quadratic term would be negative, reflecting the fact that the output increases at a decreasing rate. It should be noted that having x and x^2 in the regression introduces a certain degree of collinearity.

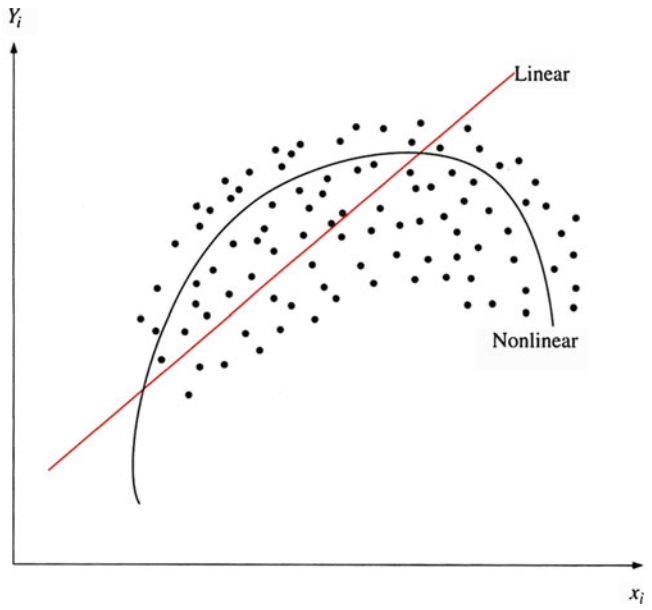


Fig. 16.14 Linear curve versus nonlinear curve

Example 16.5 A Nonlinear Market Model. Suppose the following regression model is run:

$$R_{i,t} = a + b_1 R_{m,t} + b_2 R_{m,t}^2 + e_{i,t} \tag{16.17}$$

where $R_{i,t}$ is the rate of return on Johnson & Johnson stock during 1970 to 2009 and $R_{m,t}$ is the rate of return on the market index (the S&P 500). The estimated results we get when we add a quadratic term for $R_{m,t}$ are

$$\hat{R}_{i,t} = .0233 + 0.865 R_{m,t} - 1.928 R_{m,t}^2, \qquad R^2 = .124$$
$$(0.34) \quad (2.29) \quad (-0.93)$$

where t-values are in parentheses. From Table A4 in Appendix A, by interpolation, we find that $t_{.025,38} = 2.025$. Because only 2.29 is larger than 2.025, we conclude that estimated b_1 is significantly different from 0 at $\alpha = .05$, but estimated b_2 is insignificant. These results imply that $R_{m,t}^2$ should not be included in the regression because the t-value associated with this quadratic term is statistically insignificant at $\alpha = .05$.

16.6.2 The Log-Linear and the Log–Log-Linear Model

A common transformed linear model involves the e-based logarithmic transformation of variables such as the one shown in Eq. 16.18[4]:

$$\log_e Y_i = \alpha + \beta_1(\log_e x_{1i}) + \beta_2(\log_e x_{2i}) + \cdots + \beta_n(\log_e x_{ni}) + \epsilon_i \tag{16.18}$$

The *log–log linear model* is a linear model with logarithmic transformation made on both dependent and independent variables. If only the dependent variable is being lognormally transformed, then we call this linear model a *log-linear model*. As in the quadratic case, a visual inspection of the data may help determine whether a model should be specified in a log form.

The coefficients of a log–log-linear model are elasticity coefficients, which give the percentage change in the dependent variable that is due to a 1 % change in the independent variable. For example, suppose the demand relationship

$$Q = aP_1^b X^c P_2^d$$

has been specified, where Q is quantity purchased, P_1 is price, X is income, P_2 is the price of a competing good, and a, b, c, and d are parameters. Then b, c, and d are elasticities of P_1, X, and P_2, respectively.[5]

Example 16.6 The Relationship Between Cylinder Volume and Miles per Gallon. To study the relationship between cylinder volume and miles per gallon, we use the 1986 EPA mileage guide which gives the engine size and estimated city miles per gallon ratings for 11 gasoline-fueled subcompact and compact cars. That data, as given in Table 16.5, was used to estimate the following regression relationships:

$$y_i = a + bx_i + e_i \tag{16.19a}$$

$$\log_e y_i = a' + b' \log_e x_i + e'_i \tag{16.19b}$$

[4] Equation 16.18 is obtained by taking the logarithmic transformation of a model of the equation $Y_i = \alpha_0 x_{1t}^{\beta_1} x_{2t}^{\beta_2} \cdots x_{nt}^{\beta_n}$ and letting $\alpha = \log \alpha_0$.

[5] For example, the elasticity coefficient, e_p, is defined as $(dQ/dP_1)(P_1/Q)$. The first derivative is $dQ/dP_1 = abP_1^{b-1} X^c P_2^d$

Substituting Q and this equation into the definition of e_p, we obtain

$$e_p = abP_1^{b-1} X^c P_2^d \left(\frac{P_1}{aP_1^b X^c P_2^d} \right)$$

$$= b$$

Table 16.5 Cylinder volume and miles per gallon for 11 different kinds of cars

Car	Cylinder volume, x	Miles per gallon, y
VW Golf	97	37
Chevy cavalier	173	19
Plymouth horizon	97	31
Pontiac firebird	151	23
Corvette	350	17
Honda accord	119	27
Dodge omni	97	31
Renault alliance	85	35
Olds firenza	173	19
Nissan sentra	97	31
Ford escort	114	32

Source: 1986 Gas Mileage Guide, EPA Fuel Economy Estimates, U.S. Dept. of Energy. Wards Automotive Yearbook (1986)

where

y_i = miles per gallon for ith kind of car
X_i = cylinder volume for ith kind of car

Equation 16.19a is a linear model, and Eq. 16.19b is a log–log-linear model that is similar to Eq. 16.18. MINITAB regression outputs for Eqs. 16.19a and 16.19b are presented in Figs. 16.15 and 16.16, respectively. From these outputs, the estimates regression lines are

$$\begin{aligned} \hat{y}_i &= 37.677 - .07241x_i \qquad R^2 = .634 \\ &\quad (12.95) \quad (-3.95) \\ \log\ \hat{y}_i &= 6.2020 - .60133 \log\ x_i \quad R^2 = .841 \\ &\quad (14.61) \quad (-6.90) \end{aligned}$$

t-values are in parentheses.

From Table A4 in Appendix A, we find $t_{.005,9} = 3.250$. Because all absolute t-values are larger than 3.250, all estimated parameters are significantly different from 0 at $\alpha = .01$. The bottom portion of Fig. 16.15 presents the scatter diagram of residuals against the independent variable x, which shows that there are some patterns and, therefore, heteroscedasticity in the residuals. The scatter diagram presented in the bottom portion of Fig. 16.16, however, shows that there is no heteroscedasticity in the residuals of log–log-linear regression. These results indicate that the logarithmic transformation will alleviate the heteroscedasticity among residuals. Also note that the R^2 of the log–log-linear model is .841, which is larger than that of the linear model (.634). Finally, $\hat{b}' = -.60133$ implies that with a 1 % increase in cylinder volume, miles per gallon will decrease by .60133 %.

```
MTB > READ C1 C2
DATA> 97 37
DATA> 173 19
DATA> 97 31
DATA> 151 23
DATA> 350 17
DATA> 119 27
DATA> 97 31
DATA> 85 35
DATA> 173 19
DATA> 97 31
DATA> 114 32
DATA> END
     11 rows read.
MTB > BRIEF 2
MTB > REGRESS C2 1 C1;
SUBC> RESIDUAL C3;
SUBC> DW.
```

Regression Analysis

```
The regression equation is
C2 = 37.7 - 0.0724 C1

Predictor       Coef        StDev           T        P
Constant      37.677        2.909       12.95    0.000
C1          -0.07241      0.01833       -3.95    0.003

S = 4.410       R-Sq = 63.4%     R-Sq(adj) = 59.4%

Analysis of Variance

Source       DF           SS           MS          F        P
Regression    1       303.67       303.67      15.61    0.003
Error         9       175.06        19.45
Total        10       478.73

Unusual Observations
Obs        C1          C2         Fit  StDev Fit   Residual    St Resid
  5       350       17.00       12.33        4.05       4.67       2.68RX

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 2.78

MTB > GSTD
* NOTE  * Standard Graphics are enabled.
          Professional Graphics are disabled.
          Use the GPRO command to enable Professional Graphics.


MTB > PLOT C3 C1
```

Fig. 16.15 (continued)

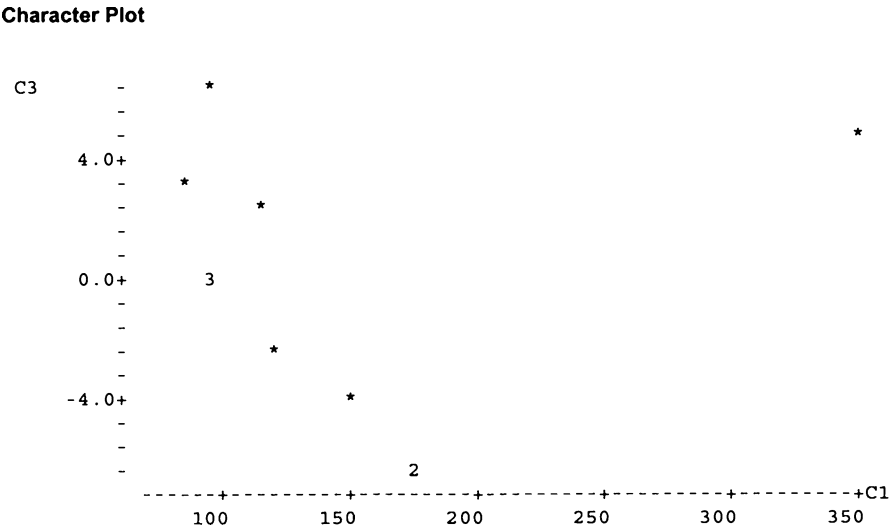


Fig. 16.15 MINITAB output of Eq. 16.19a

16.7 Lagged Dependent Variables (Optional)

In all the models we have discussed, the dependent variable was a function of independent variables in period t. However, for time-series data, we often want to lag the dependent variable by one period to estimate the effect on the variable from a previous period. The model is

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + \gamma Y_{t-1} + \epsilon_t \tag{16.20}$$

Here, the dependent variable is a function of the X's and of the dependent variable lagged one period.

A regression can be run on the data to estimate the coefficients of Eq. 16.20. However, our interpretation of these estimated coefficients must be modified to take into account the long-run effect. The short-run (current) effect is that a 1-unit increase in X_k leads to a β_k-unit increase in Y. This is the usual interpretation of regression coefficients. The long-run effect of regression coefficients is

$$\beta_i^L = \frac{\beta_i}{1-\gamma}, \quad i = 1, 2, \ldots, k \tag{16.21}$$

where β_i^L represents the long-run coefficient that takes the lagged effect into account and γ is the coefficient associated with the lagged dependent variable as defined in Eq. 16.20. In a moment, we will offer two examples to show how Eq. 16.21 is calculated.

```
MTB > LET C4=LOGE(C1)
MTB > LET C5=LOGE(C2)
MTB > PRINT C1 C2 C4 C5
```

Data Display

```
 Row    C1   C2       C4       C5

   1    97   37  4.57471  3.61092
   2   173   19  5.15329  2.94444
   3    97   31  4.57471  3.43399
   4   151   23  5.01728  3.13549
   5   350   17  5.85793  2.83321
   6   119   27  4.77912  3.29584
   7    97   31  4.57471  3.43399
   8    85   35  4.44265  3.55535
   9   173   19  5.15329  2.94444
  10    97   31  4.57471  3.43399
  11   114   32  4.73620  3.46574

MTB > REGRESS C5 1 C4;
SUBC> RESIDUAL C6;
SUBC> DW.
```

Regression Analysis

```
The regression equation is
C5 = 6.20 - 0.601 C4

Predictor       Coef       StDev          T        P
Constant      6.2020      0.4246      14.61    0.000
C4          -0.60133     0.08711      -6.90    0.000

S = 0.1140      R-Sq = 84.1%     R-Sq(adj) = 82.4%

Analysis of Variance

Source        DF          SS          MS         F        P
Regression     1     0.61985     0.61985     47.66    0.000
Error          9     0.11706     0.01301
Total         10     0.73691

Unusual Observations
Obs        C4         C5         Fit   StDev Fit    Residual    St Resid
  5      5.86     2.8332     2.6794      0.0936      0.1538       2.36RX

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 2.31


MTB > PLOT C6 C4
```

Fig. 16.16 MINITAB Output of Eq. 16.19b

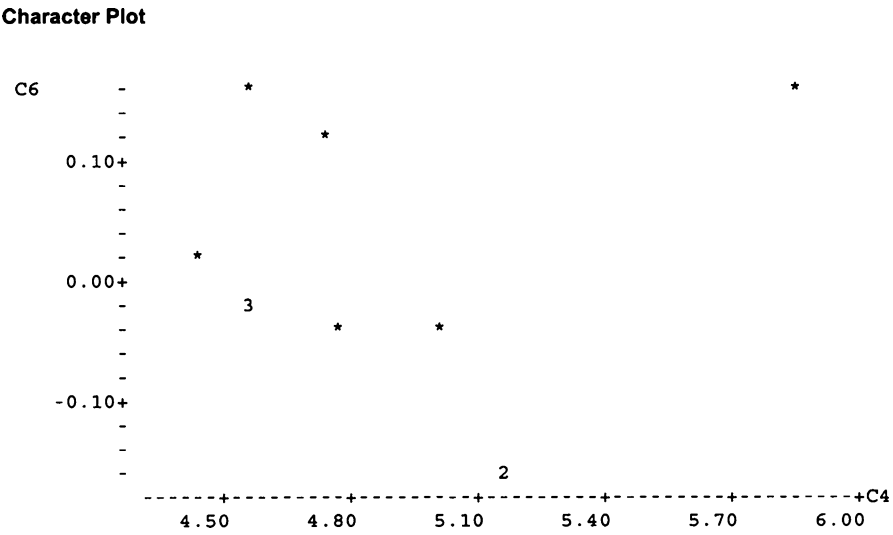


Fig. 16.16 (continued)

When a lagged dependent variable is used in the regression, the Durbin–Watson statistic is not a reliable indicator of autocorrelation. Another statistic—the *Durbin H*—is used instead. This statistic is

$$\mathrm{DH} = (1 - d/2)\sqrt{\frac{n}{1 - n\hat{V}(\hat{\gamma})}} \tag{16.22}$$

where d is the Durbin–Watson statistic defined in Eq. 16.11 and $\hat{V}(\hat{\gamma})$ is the least-squares estimate of the variance of the coefficient of the lagged variable. Under the null hypothesis, H is normally distributed with mean zero and variance 1. Therefore, the Z statistic of normal distribution can be used to do the test.

Example 16.7 The Relationship Between Dividend per Share and Earnings per Share. MINITAB outputs of two regressions of Eq. 16.23 for JNJ and MRK for period 1990–2009 are presented in Figs. 16.17 and 16.18, respectively:

$$\mathrm{DPS}_{i,t} = \alpha + \beta \mathrm{EPS}_{i,t} + \gamma \mathrm{DPS}_{i,t-1} + \epsilon_{i,t} \tag{16.23}$$

Equation 16.23 is obtained by adding a variable for lagged dividend per share (DPS_{t-1}) to the right-hand side of Eq. 16.13.

The results shown in Figs. 16.17 and 16.18 indicate that t statistics of the γ coefficient for JNJ and MRK are 3.60 and 0.90, respectively. Therefore, lagged

Regression Analysis: DPS(JNJ) versus EPS(JNJ), lagDPS(JNJ)

The regression equation is
DPS(JNJ) = -0.161 + 0.323 EPS(JNJ) + 0.304 lagDPS(JNJ)

Predictor	Coef	SE Coef	T	P	VIF
Constant	-0.16118	0.09389	-1.72	0.104	
EPS(JNJ)	0.32326	0.02918	11.08	0.000	1.340
lagDPS(JNJ)	0.30374	0.08443	3.60	0.002	1.340

S = 0.0955327 R-Sq = 93.3% R-Sq(adj) = 92.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	2.1501	1.0750	117.79	0.000
Residual Error	17	0.1552	0.0091		
Total	19	2.3052			

Source	DF	Seq SS
EPS(JNJ)	1	2.0320
lagDPS(JNJ)	1	0.1181

Obs	EPS(JNJ)	DPS(JNJ)	Fit	SE Fit	Residual	St Resid
1	3.38	1.2877	1.2653	0.0240	0.0223	0.24
2	4.30	1.5084	1.6199	0.0386	-0.1116	-1.28
3	1.54	0.8769	0.7948	0.0698	0.0821	1.26 X
4	2.71	1.0003	0.9812	0.0278	0.0191	0.21
5	3.08	1.1155	1.1383	0.0239	-0.0227	-0.25
6	3.65	1.2546	1.3575	0.0280	-0.1030	-1.13
7	2.12	0.7157	0.9052	0.0401	-0.1895	-2.19R
8	2.41	0.8300	0.8352	0.0365	-0.0053	-0.06
9	2.23	0.9514	0.8118	0.0321	0.1397	1.55
10	2.94	1.0429	1.0782	0.0250	-0.0353	-0.38
11	3.39	1.2163	1.2514	0.0259	-0.0351	-0.38
12	1.83	0.6605	0.7998	0.0453	-0.1394	-1.66
13	2.16	0.7796	0.7377	0.0402	0.0419	0.48
14	2.39	0.9129	0.8482	0.0331	0.0647	0.72
15	2.83	1.0824	1.0309	0.0263	0.0515	0.56
16	3.46	1.2591	1.2861	0.0257	-0.0270	-0.29
17	3.73	1.4411	1.4270	0.0276	0.0141	0.15
18	3.63	1.6044	1.4500	0.0319	0.1545	1.72
19	4.57	1.7718	1.8034	0.0478	-0.0317	-0.38
20	4.40	1.9099	1.7993	0.0535	0.1106	1.40

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
Durbin-Watson statistic = 1.85326

Fig. 16.17 MINITAB output of Eq. 16.23 for JNJ

Regression Analysis: DPS(MRK) versus EPS(MRK), lagDPS(MRK)

The regression equation is
DPS(MRK) = 0.630 + 0.203 EPS(MRK) + 0.146 lagDPS(MRK)

Predictor	Coef	SE Coef	T	P	VIF
Constant	0.6297	0.2605	2.42	0.027	
EPS(MRK)	0.20315	0.04696	4.33	0.000	1.039
lagDPS(MRK)	0.1463	0.1620	0.90	0.379	1.039

S = 0.233882 R-Sq = 56.2% R-Sq(adj) = 51.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1.19487	0.59743	10.92	0.001
Residual Error	17	0.92991	0.05470		
Total	19	2.12478			

Source	DF	Seq SS
EPS(MRK)	1	1.15028
lagDPS(MRK)	1	0.04459

Obs	EPS(MRK)	DPS(MRK)	Fit	SE Fit	Residual	St Resid
1	4.51	1.9967	1.7950	0.0869	0.2018	0.93
2	5.39	2.3388	2.0168	0.1339	0.3220	1.68
3	1.70	0.9468	1.3172	0.1728	-0.3704	-2.35RX
4	1.86	1.0628	1.1461	0.1078	-0.0833	-0.40
5	2.35	1.1465	1.2626	0.0872	-0.1161	-0.54
6	2.63	1.2430	1.3317	0.0753	-0.0887	-0.40
7	3.12	1.4403	1.4454	0.0650	-0.0051	-0.02
8	3.74	1.6966	1.6002	0.0615	0.0964	0.43
9	4.30	1.9277	1.7514	0.0802	0.1763	0.80
10	2.45	1.0934	1.4094	0.0987	-0.3160	-1.49
11	2.90	1.2348	1.3788	0.0808	-0.1440	-0.66
12	3.14	1.3590	1.4482	0.0660	-0.0892	-0.40
13	3.14	1.4072	1.4664	0.0561	-0.0592	-0.26
14	3.03	1.4490	1.4511	0.0536	-0.0021	-0.01
15	2.61	1.4953	1.3719	0.0568	0.1234	0.54
16	2.10	1.5173	1.2750	0.0704	0.2423	1.09
17	2.03	1.5170	1.2640	0.0733	0.2529	1.14
18	1.49	1.5097	1.1543	0.0927	0.3554	1.66
19	3.64	1.5151	1.5900	0.0581	-0.0749	-0.33
20	5.68	1.5836	2.0052	0.1315	-0.4217	-2.18R

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.

Durbin-Watson statistic = 1.31920

Fig. 16.18 MINITAB output of Eq. 16.23 for MRK

dividend per share is important in explaining dividend per share in period t for only JNJ.

Long-run coefficients (accumulated effect over all future periods) associated with EPS in terms of Eq. 16.21 are calculated as follows:

$$\frac{.323}{1-.304} = .4641 \text{ long-run coefficient for JNJ}$$
$$\frac{.203}{1-.146} = .2377 \text{ long-run coefficient for MRK}$$

These long-run coefficients imply that a $1.00 increase in EPS will spell a total increase of $.4641 and $.2377 in DPS for JNJ and MRK, respectively. Total dividend increases are much higher than short-run (current) increases of $.323 and $.203 for JNJ and MRK.

Example 16.8 Time Aggregation and the Estimation of the Market Model. In application 14.2 of Chap. 14, we use market model to investigate the relationship between rates of return for individual securities and market rates of return. The coefficient of market rates of return in market model is called *Beta* coefficient, which is used to determine the degree of nondiversifiable risk of a firm. Cartwright and Lee (1987)[6] used data for heavily and lightly traded firms to evaluate the effects of temporal aggregation on beta estimates. They indicated the importance of price adjustment delays in the trading process and found that temporal aggregation has important effects on the market model. A regression of Eq. 16.24 is used to find the impact coefficient of temporal aggregation:

$$R_{i,t} = \alpha_i + \beta_{i1}R_{i,t-1} + \beta_{i2}R_{m,t} + \epsilon_t \tag{16.24}$$

where $R_{i,t}$ is the rate of return for the ith firm in period t, $R_{i,t-1}$ is the lagged rate of return for the ith firm, $R_{m,t}$ is the market rate of return, and $e_{i,t}$ is the error term. The *Beta* coefficient β_{i2} is the traditional risk measure, and $\beta_{i2}/(1\text{-}\beta_{i1})$ is the long-run coefficient to represent the impact of temporal aggregation on systematic risk measure.

Here, we use the annual rates of return in 1970–2009 for Johnson & Johnson and Merck as the examples. MINITAB outputs of two regressions of Eq. 16.24 for JNJ and MRK are presented in Figs. 16.19 and 16.20, respectively.

The results shown in Figs. 16.19 and 16.20 indicate that t statistics of β_{i1} coefficient for JNJ and MRK are −3.38 and −2.19, respectively. Therefore, lagged rate of return is important in explaining rate of return in period t for both JNJ and MRK.

[6] See Cartwright, Lee.: Time aggregation and the estimation of the market model: Empirical evidence. J. Bus. Econ. Stat. **5**(1), 131–143 (1987)

Long-run coefficients associated with annual market rates of return in terms of Eq. 16.24 are calculated as follows:

$$\frac{.9433}{1-(-0.4514)} = .6499 \text{ long-run coefficient for JNJ}$$

$$\frac{.9542}{1-(-0.3228)} = .7213 \text{ long-run coefficient for MRK}$$

Data Display

Row	Year	JNJ	MRK	S&P	JNJ(lag)	MRK(lag)
1	1970	-0.681455	-0.105661	-0.149428	0.698039	0.278422
2	1971	0.735557	0.274854	0.181086	-0.681455	-0.105661
3	1972	0.329385	-0.272215	0.110998	0.735557	0.274854
4	1973	-0.132041	-0.080191	-0.016209	0.329385	-0.272215
5	1974	-0.276278	-0.160629	-0.228800	-0.132041	-0.080191
6	1975	0.120214	0.064286	0.039952	-0.276278	-0.160629
7	1976	-0.119259	0.004346	0.183960	0.120214	0.064286
8	1977	0.001873	-0.162669	-0.037349	-0.119259	0.004346
9	1978	-0.017184	0.249828	-0.022200	0.001873	-0.162669
10	1979	0.101555	0.097778	0.072797	-0.017184	0.249828
11	1980	0.286313	0.205990	0.153092	0.101555	0.097778
12	1981	-0.619442	0.031377	0.078043	0.286313	0.205990
13	1982	0.362091	0.031582	-0.065131	-0.619442	0.031377
14	1983	-0.155059	0.101627	0.339988	0.362091	0.031582
15	1984	-0.087818	0.074488	0.000312	-0.155059	0.101627
16	1985	0.491250	0.493088	0.164402	-0.087818	0.074488
17	1986	0.272454	-0.080899	0.264933	0.491250	0.493088
18	1987	0.165022	0.300894	0.213633	0.272454	-0.080899
19	1988	0.162139	-0.627023	-0.073354	0.165022	0.300894
20	1989	-0.289582	0.371471	0.214643	0.162139	-0.627023
21	1990	0.230108	0.185441	0.036396	-0.289582	0.371471
22	1991	0.616842	0.878595	0.124301	0.230108	0.185441
23	1992	-0.551293	-0.733803	0.105162	0.616842	0.878595
24	1993	-0.091578	-0.182990	0.085799	-0.551293	-0.733803
25	1994	0.244915	0.142442	0.019960	-0.091578	-0.182990
26	1995	0.584558	0.753915	0.176578	0.244915	0.142442
27	1996	-0.409758	0.235280	0.237724	0.584558	0.753915
28	1997	0.340804	0.352548	0.302655	-0.409758	0.235280
29	1998	0.287688	0.409696	0.242801	0.340804	0.352548
30	1999	0.124207	-0.537078	0.222782	0.287688	0.409696
31	2000	0.139719	0.411867	0.075256	0.124207	-0.537078
32	2001	-0.431191	-0.357447	-0.163282	0.139719	0.411867
33	2002	-0.078010	-0.013313	-0.167680	-0.431191	-0.357447
34	2003	-0.021172	-0.158294	-0.028885	-0.078010	-0.013313
35	2004	0.248595	-0.271964	0.171379	-0.021172	-0.158294
36	2005	-0.032496	0.036942	0.067731	0.248595	-0.271964
37	2006	0.122480	0.418327	0.085510	-0.032496	0.036942
38	2007	0.034602	0.367425	0.127230	0.122480	0.418327
39	2008	-0.076435	-0.450781	-0.174081	0.034602	0.367425
40	2009	0.108473	0.254065	-0.222935	-0.076435	-0.450781

Fig. 16.19 MINITAB output of Eq. 16.24 for JNJ

Regression Analysis: JNJ versus JNJ(lag), S&P

The regression equation is
JNJ = 0.0159 - 0.451 JNJ(lag) + 0.943 S&P

Predictor	Coef	SE Coef	T	P
Constant	0.01585	0.04818	0.33	0.744
JNJ(lag)	-0.4514	0.1334	-3.38	0.002
S&P	0.9433	0.3078	3.06	0.004

S = 0.274393 R-Sq = 31.4% R-Sq(adj) = 27.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1.27688	0.63844	8.48	0.001
Residual Error	37	2.78578	0.07529		
Total	39	4.06267			

Source	DF	Seq SS
JNJ(lag)	1	0.56965
S&P	1	0.70723

Obs	JNJ(lag)	JNJ	Fit	SE Fit	Residual	St Resid
1	0.698	-0.6815	-0.4402	0.1270	-0.2413	-0.99
2	-0.681	0.7356	0.4943	0.1208	0.2413	0.98
3	0.736	0.3294	-0.2115	0.0975	0.5408	2.11R
4	0.329	-0.1320	-0.1481	0.0650	0.0161	0.06
5	-0.132	-0.2763	-0.1404	0.0992	-0.1359	-0.53
6	-0.276	0.1202	0.1782	0.0621	-0.0580	-0.22
7	0.120	-0.1193	0.1351	0.0554	-0.2544	-0.95
8	-0.119	0.0019	0.0345	0.0564	-0.0326	-0.12
9	0.002	-0.0172	-0.0059	0.0513	-0.0113	-0.04
10	-0.017	0.1016	0.0923	0.0449	0.0093	0.03
11	0.102	0.2863	0.1144	0.0502	0.1719	0.64
12	0.286	-0.6194	-0.0398	0.0521	-0.5797	-2.15R
13	-0.619	0.3621	0.2340	0.1010	0.1281	0.50
14	0.362	-0.1551	0.1731	0.0943	-0.3282	-1.27
15	-0.155	-0.0878	0.0861	0.0539	-0.1740	-0.65
16	-0.088	0.4912	0.2106	0.0587	0.2807	1.05
17	0.491	0.2725	0.0440	0.0847	0.2284	0.88
18	0.272	0.1650	0.0944	0.0637	0.0706	0.26
19	0.165	0.1621	-0.1278	0.0651	0.2900	1.09
20	0.162	-0.2896	0.1451	0.0616	-0.4347	-1.63
21	-0.290	0.2301	0.1809	0.0633	0.0492	0.18
22	0.230	0.6168	0.0292	0.0498	0.5876	2.18R
23	0.617	-0.5513	-0.1634	0.0839	-0.3879	-1.48
24	-0.551	-0.0916	0.3456	0.0943	-0.4372	-1.70
25	-0.092	0.2449	0.0760	0.0490	0.1689	0.63
26	0.245	0.5846	0.0719	0.0565	0.5127	1.91
27	0.585	-0.4098	-0.0238	0.0878	-0.3860	-1.48
28	-0.410	0.3408	0.4863	0.1149	-0.1455	-0.58

Fig. 19.19 (continued)

29	0.341	0.2877	0.0911	0.0720	0.1966	0.74
30	0.288	0.1242	0.0962	0.0660	0.0281	0.11
31	0.124	0.1397	0.0308	0.0440	0.1089	0.40
32	0.140	-0.4312	-0.2012	0.0861	-0.2300	-0.88
33	-0.431	-0.0780	0.0523	0.0967	-0.1303	-0.51
34	-0.078	-0.0212	0.0238	0.0537	-0.0450	-0.17
35	-0.021	0.2486	0.1871	0.0564	0.0615	0.23
36	0.249	-0.0325	-0.0325	0.0498	-0.0000	-0.00
37	-0.032	0.1225	0.1112	0.0460	0.0113	0.04
38	0.122	0.0346	0.0806	0.0469	-0.0460	-0.17
39	0.035	-0.0764	-0.1640	0.0857	0.0875	0.34
40	-0.076	0.1085	-0.1599	0.0975	0.2684	1.05

R denotes an observation with a large standardized residual.
Durbin-Watson statistic = 2.16451

Fig. 16.19 (continued)

These long-run coefficients imply that a one-unit change increase in annual market rates of return will spell a total change increase of .6499 and .7213 in annual rates of return for JNJ and MRK, respectively.

Example 16.9 Consumption Function Analysis. In this example, a consumption function is specified with a lagged dependent variable. A consumption function measures the change in consumption that is attributable to a 1-unit change in income. If the slope term in the regression for income is .75, for example, individuals tend to spend 75 cents out of every additional dollar earned. (The slope term is called the *marginal propensity to consume*, or MPC.) A regression of Eq. 16.25 is run with personal consumption (C_t) in the United States from 1962 to 2009 as the dependent variable and with disposable income (DI_t) and personal consumption lagged one period as the independent variables:

$$C_t = \alpha + \beta_1 DI_t + \beta_2 C_{t-1} + \epsilon_t \tag{16.25}$$

The data used to run this regression are listed in Table 16.6, and the results are presented in Table 16.7. The critical t-value used to do the test is $t_{.005,40} = 2.704$.

With a t-value of -2.72, the constant is statistically different from zero. Disposable income has a coefficient of .487. This implies that individuals will consume about 48.7 cents out of every additional dollar in income. At 5.71, the t statistic is significant at every level of significance. The lagged consumption variable is also highly significant; it has a coefficient of .501. If disposable income increases by 1 unit in the current period, the expected increase in consumption is .487 in the current period, is $.487 \times .501 = .244$ the next year, is $(.501)^2 \times .487 = .122$ two years later, and so on. The total increase on all future consumption in terms of Eq. 16.21 is .487/(1–.501) = .976. Note that the long-run coefficient, .976, is much larger than the short-run coefficient, .487.

Substituting $n = 48, d = .632$, and $V(\hat{\beta}_2) = (.090)^2 = .0081$ into Eq. 16.22, we obtain

$$\text{DH} = \left(1 - \frac{.632}{2}\right)\sqrt{\frac{48}{1 - 48(.0081)}} = 6.0616$$

Regression Analysis: MRK versus MRK(lag), S&P

The regression equation is
MRK = 0.0190 -0.323 MRK(lag) + 0.954 S&P

Predictor	Coef	SE Coef	T	P
Constant	0.01904	0.05497	0.35	0.731
MRK(lag)	-0.3228	0.1475	-2.19	0.035
S&P	0.9542	0.3473	2.75	0.009

S = 0.312749 R-Sq = 22.1% R-Sq(adj) = 17.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1.02389	0.51195	5.23	0.010
Residual Error	37	3.61904	0.09781		
Total	39	4.64293			

Source	DF	Seq SS
MRK(lag)	1	0.28568
S&P	1	0.73822

Obs	MRK(lag)	MRK	Fit	SE Fit	Residual	St Resid
1	0.278	-0.1057	-0.2134	0.1004	0.1078	0.36
2	-0.106	0.2749	0.2259	0.0704	0.0489	0.16
3	0.275	-0.2722	0.0362	0.0588	-0.3084	-1.00
4	-0.272	-0.0802	0.0914	0.0724	-0.1716	-0.56
5	-0.080	-0.1606	-0.1734	0.1130	0.0128	0.04
6	-0.161	0.0643	0.1090	0.0594	-0.0447	-0.15
7	0.064	0.0043	0.1738	0.0636	-0.1695	-0.55
8	0.004	-0.1627	-0.0180	0.0613	-0.1447	-0.47
9	-0.163	0.2498	0.0504	0.0646	0.1995	0.65
10	0.250	0.0978	0.0078	0.0564	0.0899	0.29
11	0.098	0.2060	0.1336	0.0572	0.0724	0.24
12	0.206	0.0314	0.0270	0.0535	0.0044	0.01
13	0.031	0.0316	-0.0532	0.0674	0.0848	0.28
14	0.032	0.1016	0.3333	0.1073	-0.2316	-0.79
15	0.102	0.0745	-0.0135	0.0556	0.0880	0.29
16	0.074	0.4931	0.1519	0.0594	0.3412	1.11
17	0.493	-0.0809	0.1127	0.0974	-0.1936	-0.65
18	-0.081	0.3009	0.2490	0.0764	0.0519	0.17
19	0.301	-0.6270	-0.1481	0.0821	-0.4789	-1.59

Fig. 16.20 MINITAB output of Eq. 16.24 for MRK

20	-0.627	0.3715	0.4263	0.1317	-0.0548	-0.19
21	0.371	0.1854	-0.0662	0.0694	0.2516	0.83
22	0.185	0.8786	0.0778	0.0549	0.8008	2.60R
23	0.879	-0.7338	-0.1643	0.1283	-0.5695	-2.00
24	-0.734	-0.1830	0.3378	0.1289	-0.5208	-1.83
25	-0.183	0.1424	0.0972	0.0619	0.0453	0.15
26	0.142	0.7539	0.1415	0.0618	0.6124	2.00
27	0.754	0.2353	0.0025	0.1184	0.2328	0.80
28	0.235	0.3525	0.2319	0.0945	0.1207	0.40
29	0.353	0.4097	0.1369	0.0834	0.2728	0.90
30	0.410	-0.5371	0.0994	0.0830	-0.6364	-2.11R
31	-0.537	0.4119	0.2642	0.1020	0.1476	0.50
32	0.412	-0.3574	-0.2697	0.1144	-0.0877	-0.30
33	-0.357	-0.0133	-0.0256	0.1057	0.0123	0.04
34	-0.013	-0.1583	-0.0042	0.0599	-0.1541	-0.50
35	-0.158	-0.2720	0.2337	0.0723	-0.5056	-1.66
36	-0.272	0.0369	0.1715	0.0700	-0.1345	-0.44
37	0.037	0.4183	0.0887	0.0500	0.3296	1.07
38	0.418	0.3674	0.0054	0.0721	0.3620	1.19
39	0.367	-0.4508	-0.2657	0.1138	-0.1851	-0.64
40	-0.451	0.2541	-0.0482	0.1250	0.3022	1.05

R denotes an observation with a large standardized residual.
Durbin-Watson statistic = 1.95251

Fig. 16.20 (continued)

Using Table A3 in Appendix A, we find that DWH = 6.0616 is larger than $z = 3$ ($\alpha = .0013$). Hence, there is autocorrelation associated with this consumption function.

To adjust for the impact of autocorrelation, we can use a modified regression model to estimate the consumption function. It is

$$C_t - \hat{\rho}C_{t-1} = \alpha(1 - \hat{\rho}) + \beta_1(\mathrm{DI}_t - \hat{\rho}\mathrm{DI}_{t-1}) + \beta_2(C_{t-1} - \hat{\rho}C_{t-2}) + \epsilon'_t \quad (16.25a)$$

where $\epsilon'_t = \epsilon_t - \hat{\rho}\epsilon_{t-1}$

Plugging the data listed in Table 16.6 and $\hat{\rho} = .684$ into Eq. 16.25a yields the results presented in Table 16.8. These results are more appropriate for null hypothesis testing than are those indicated in Table 16.7.

16.8 Dummy Variables

So far, we have used data that could take on any number of values. In this section, we will examine an independent variable that can take on either of just two values: 1 and 0. This binary variable is called a *dummy variable*, and it enables us to include information that is not quantitative. For example, a regression that models individuals' income might include a dummy variable for sex of the worker. The independent dummy variable for sex could take on the value 1 for a male worker

Table 16.6 Personal consumption and disposable income (in billions of 2005 dollars)

Year	C_t	DI_t	C_{t-1}
1961	1,821.2	2,030.8	1,784.4
1962	1,911.2	2,129.6	1,821.2
1963	1,989.9	2,209.5	1,911.2
1964	2,108.4	2,368.7	1,989.9
1965	2,241.8	2,514.7	2,108.4
1966	2,369.0	2,647.3	2,241.8
1967	2,440.0	2,763.5	2,369.0
1968	2,580.7	2,889.2	2,440.0
1969	2,677.4	2,981.4	2,580.7
1970	2,740.2	3,108.8	2,677.4
1971	2,844.6	3,249.1	2,740.2
1972	3,019.5	3,406.6	2,844.6
1973	3,169.1	3,638.2	3,019.5
1974	3,142.8	3,610.2	3,169.1
1975	3,214.1	3,691.3	3,142.8
1976	3,393.1	3,838.3	3,214.1
1977	3,535.9	3,970.7	3,393.1
1978	3,691.8	4,156.5	3,535.9
1979	3,779.5	4,253.8	3,691.8
1980	3,766.2	4,295.6	3,779.5
1981	3,823.3	4,410.0	3,766.2
1982	3,876.7	4,506.5	3,823.3
1983	4,098.3	4,655.7	3,876.7
1984	4,315.6	4,989.1	4,098.3
1985	4,540.4	5,144.8	4,315.6
1986	4,724.5	5,315.0	4,540.4
1987	4,870.3	5,402.4	4,724.5
1988	5,066.6	5,635.6	4,870.3
1989	5,209.9	5,785.1	5,066.6
1990	5,316.2	5,896.3	5,209.9
1991	5,324.2	5,945.9	5,316.2
1992	5,505.7	6,155.3	5,324.2
1993	5,701.2	6,258.2	5,505.7
1994	5,918.9	6,459.0	5,701.2
1995	6,079.0	6,651.6	5,918.9
1996	6,291.2	6,870.9	6,079.0
1997	6,523.4	7,113.5	6,291.2
1998	6,865.5	7,538.8	6,523.4
1999	7,240.9	7,766.7	6,865.5
2000	7,608.1	8,161.5	7,240.9
2001	7,813.9	8,360.1	7,608.1
2002	8,021.9	8,637.1	7,813.9
2003	8,247.6	8,853.9	8,021.9
2004	8,532.7	9,155.1	8,247.6
2005	8,819.0	9,277.3	8,532.7
2006	9,073.5	9,650.7	8,819.0
2007	9,313.9	9,860.6	9,073.5
2008	9,290.9	9,911.3	9,313.9
2009	9,237.3	10,035.3	9,290.9

Table 16.7 Results of regression of Eq. 16.24

Variable	Coefficient	Standard error	t-value	p-value
Constant	−109.10	40.10	−2.72	.009
DI_t	.487	.085	5.71	.0000
C_{t-1}	.501	.090	5.54	.0000

$R^2 = .9991$
$\bar{R}^2 = .9991$
Observations 48
First-order autocorrelation $(\hat{\rho}) = .684$
DW = .632

Table 16.8 Results of regression of Eq. 16.25a

Variable	Coefficient	Standard error	t-value	p-value
Constant	−48.87	22.18	−2.20	.033
DI_t	.638	.088	7.28	.358
C_{t-1}	.337	.092	3.67	.001

$R^2 = .9956$
$\bar{R}^2 = .9954$
Observations 48
DW = 1.627

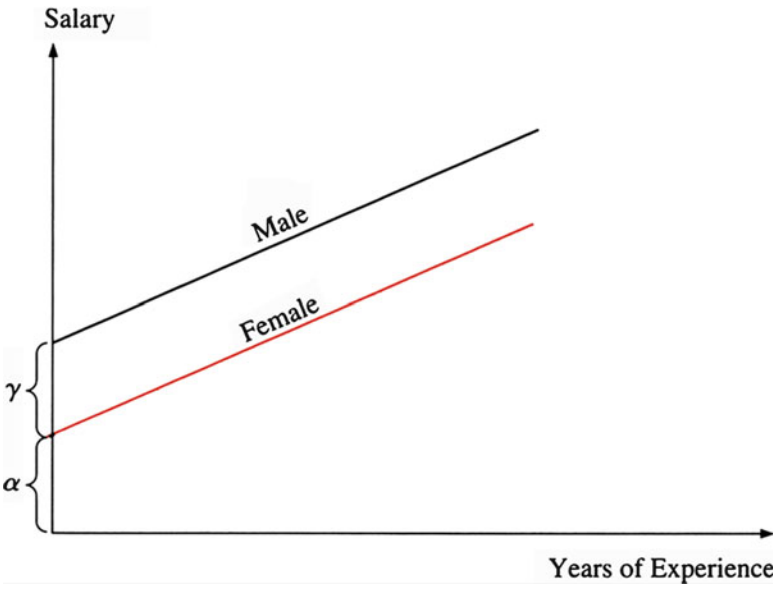


Fig. 16.21 The relationship between salary and years of experience

and the value 0 for a female worker. (The assignment of dummy variables is arbitrary; we could—and in this day and age probably should—have reversed the assignment: 1 for female, 0 for male.) The regression is

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \gamma D_{1i} + \epsilon_i \tag{16.26}$$

where the betas (β_1, β_2, ..., β_k) are the coefficients for the quantitative variables and γ is the parameter for the dummy variable:

$$\begin{aligned} D_1 &= 1 \quad \text{if the worker is a male} \\ D_1 &= 0 \quad \text{if the worker is a female} \end{aligned}$$

Dummy variables indicate whether a shift in the intercept term is attributable to the characteristic of the dummy. The dummy variable having a statistically significant coefficient of 1,214, for example, would indicate that the intercept for males is $1,214 higher than the intercept for females. Figure 16.21 plots salary and years of experience. The female intercept is given by α. If the intercept for the dummy were negative, the male intercept would be lower than the female. The dummy variable in Eq. 16.26 deals with the intercept term, not the slope term, so the dummy indicates only a shift in the intercept term. In other words, a positive and statistically significant coefficient for the dummy variable indicates that even though the salaries for males and females are affected by the same factors in the same way—that is, they have the same slope coefficients—males begin with a higher level of earnings and maintain that difference across all values for years of experience.

Example 16.10 Analysis of the Money Supply. In October 1979, the Federal Reserve's Board of Governors switched from targeting interest rates to targeting the money supply. Before this period, the Fed adhered to a policy of increasing the money supply by an amount that would keep interest rates stable; hence, it "targeted" interest rates. After October 1979, the Fed focused on increasing the money supply at a fixed rate and let interest rates seek their own equilibrium level:

$$M_{3t} = \alpha + \beta_1 \text{GNP}_t + \beta_2 \text{PRIME}_t + \gamma \text{DUM}_t + \epsilon_t \tag{16.27}$$

In this model, we investigate whether the Fed's policy change had an effect on the money supply. M_3 is the money supply, GNP is the gross national product, PRIME is the prime interest rate, and DUM is a dummy variable in which 1 equals the years 1979–1990 and 0 the years 1959–1978. A significant positive sign would indicate that the money supply was greater after the change. A negative sign would indicate that the money supply was less.

The regression of Eq. 16.27 is run using the annual data for 1959–1990 presented in Table 16.9. The regression results appear in Table 16.10.

The relationship between GNP and the money supply is extremely strong. There is a negative relationship between the money supply and the prime interest rate.

Table 16.9 GNP, prime rate, and M_3 (1959–1990)

Data Display

Row	YEAR	GNP	PRIMERt	DUMMY	GNPPRIME	M3
MTB > PRINT C1–C6						
1	59	1629.1	4.48	1	7298.4	140.0
2	60	1665.3	4.82	1	8026.7	140.7
3	61	−1706.7	4.50	1	7689.2	145.2
4	62	1799.4	4.50	1	8097.3	147.9
5	63	1873.3	4.50	1	8429.9	153.4
6	64	1973.3	4.50	1	8879.9	160.4
7	65	2087.6	4.54	1	9477.7	167.9
8	66	2208.3	5.63	1	12432.7	172.1
9	67	2271.4	5.61	1	12742.6	183.3
10	68	2365.6	6.30	1	14903.3	197.5
11	69	2423.3	7.96	1	19289.5	204.0
12	70	2416.2	7.91	1	19112.1	214.5
13	71	2484.8	5.72	1	14213.1	228.4
14	72	2608.5	5.25	1	13694.6	249.3
15	73	2744.1	8.03	1	22035.1	262.9
16	74	2729.3	10.81	1	29503.7	274.4
17	75	2695.0	7.86	1	21182.7	287.6
18	76	2826.7	6.84	1	19334.6	306.4
19	77	2958.6	6.83	1	20207.2	331.3
20	78	3115.2	9.06	1	28223.7	358.5
21	79	3192.4	12.67	0	40447.7	382.9
22	80	3187.1	15.27	0	48667.0	408.9
23	81	3248.8	18.87	0	61304.9	436.5
24	82	3166.0	14.86	0	47046.8	474.5
25	83	3279.1	10.79	0	35381.5	521.2
26	84	3501.4	12.04	0	42156.9	552.1
27	85	3618.7	9.93	0	35933.7	620.1
28	86	3717.9	8.33	0	30970.1	724.7
29	87	3845.3	8.22	0	31608.4	750.4
30	88	4016.9	9.32	0	37437.5	787.5
31	89	4117.7	10.87	0	44759.4	794.8
32	90	4155.8	10.01	0	41599.6	825.5

Table 16.10 Results of regression of Eq. 16.26

	Coefficient	*t*-value	*p*-value
Constant	−212.74	−4.84	.000
GNP	.2356	13.57	.000
PRIME	−19.066	−6.23	.000
DUM	198.41	6.60	.000
$R^2 = .924$	$F = 176.1$	DW = .17	

In addition, the dummy variable has a significant t-value at $\alpha = 1\ \%$, indicating that the money supply did increase after the Federal Reserve Board changed its policy.

16.9 Regression with Interaction Variables

The regression models specified thus far assume that there is no interaction between the independent variables. This assumption is not always realistic. In many situations, the relationship between one of the independent variables and the dependent variable is dependent on the value of another independent variable. This situation reflects *interaction*.

For example, suppose the following multiple regression model is specified:

$$\text{CROP}_t = \alpha + \beta_1 \text{RAIN}_t + \beta_2 \text{FERT}_t + \epsilon_t \tag{16.28}$$

In this model, the amount of corn a farmer produces (CROP_t) is a function of the amount of rain received in a growing season (RAIN_t) and the amount of fertilizer used (FERT_t). Note that there is no interaction in this model; the fertilizer affects the output of corn, but this effect doesn't depend on how much rain fell (see Fig. 16.22). In Fig. 16.22, fertilizer is graphed on the x axis, crop production on the y axis. The rate of increase in crop production is constant for any change in the amount of rain.

However, interaction results if more rain makes the fertilizer more productive and increases corn production. We can model this interaction between the two variables by adding an interaction term:

$$\text{CORN}_t = \alpha + \beta_1 \text{RAIN}_t + \beta_2 \text{FERT}_t + \beta_3 (\text{FERT}_t \times \text{RAIN}_t) + \epsilon_t \tag{16.29}$$

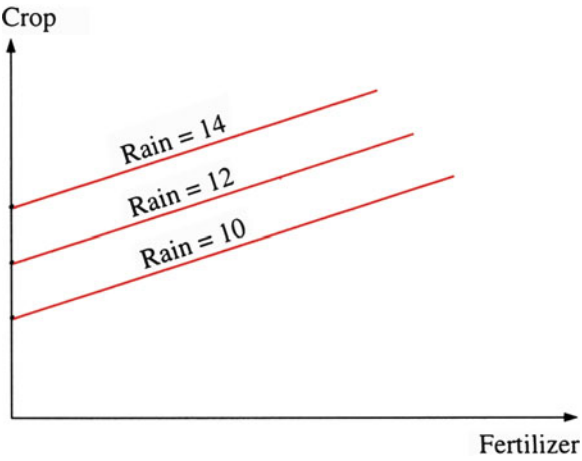


Fig. 16.22 Impact of fertilizer on output without interaction effect

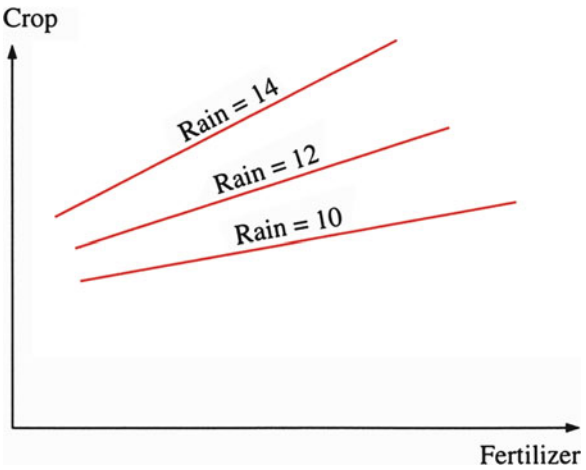


Fig. 16.23 Impact of fertilizer on output with interaction effect

To create an interaction term, we multiply the two observations whose interaction we wish to investigate. This term measures whether additional rain makes fertilizer more productive. In the model shown in Eq. 16.28, the change in the corn production that results from a change in the amount of fertilizer is given by the slope term β_2. A 1-unit change in the amount of fertilizer is associated with a β_2-unit change in crop production. In the interaction model of Eq. 16.29, the change in the corn production that is associated with a 1-unit change in the amount of fertilizer is equal to $(\beta_2 + \beta_3 \text{RAIN}_t)$. If the interaction term has a positive sign, then the rain makes the fertilizer more effective. This effectiveness is shown in Fig. 16.23, where the dependent variable is graphed on the y axis and fertilizer on the x axis. As the amount of rain increases, the slope of the line increases, indicating that fertilizer has a greater impact when more rain is present. In general, the interaction term tests whether the slope parameter for one variable changes as a function of the other variable. The t statistic is used to determine the statistical significance of the coefficient associated with the interaction term.

Example 16.11 Analysis of the Money Supply, with Interaction. Equation 16.27 with interaction can be written as

$$M_{3t} = \alpha + \beta_1 \text{GNP}_t + \beta_2 \text{PRIME}_t + \beta_3 (\text{GNP}_t)(\text{PRIME}_t) + \gamma \text{DUM}_t + \epsilon_t \quad (16.30)$$

MINITAB results for this equation are presented in Fig. 16.24. This output indicates that the t statistic associated with the interaction terra is 3.08 and that the p-value associated with the interaction term is .005. Hence, the coefficient associated with the interaction term is significantly different from 0 at $\alpha = 1$ %.

```
MTB > BRIEF 3
MTB > REGRESS 'M3' 4 'GNP' 'PRIMERt' 'DUMMY' 'GNPPRIME';
SUBC> DW.
```

Regression Analysis

```
The regression equation is
M3 = 198 + 0.132 GNP - 71.1 PRIMERt - 127 DUMMY + 0.0186 GNPPRIME

Predictor       Coef        StDev          T        P
Constant      197.93        92.18       2.15    0.041
GNP          0.13247      0.03680       3.60    0.001
PRIMERt       -71.14        17.12      -4.15    0.000
DUMMY         -126.55        35.20      -3.60    0.001
GNPPRIME    0.018589     0.006038       3.08    0.005

S = 35.31       R-Sq = 97.8%     R-Sq(adj) = 97.5%

Analysis of Variance

Source       DF          SS          MS          F        P
Regression    4     1493207      373302     299.46    0.000
Error        27       33658        1247
Total        31     1526864

Source       DF      Seq SS
GNP           1     1391385
PRIMERt       1       19320
DUMMY         1       70686
GNPPRIME      1       11815

Obs        GNP          M3         Fit   StDev Fit    Residual     St Resid
  1       1629      140.00      104.17       15.00       35.83        1.12
  2       1665      140.70       98.32       13.56       42.38        1.30
  3       1709      145.20      120.55       13.94       24.65        0.76
  4       1799      147.90      140.15       12.87        7.75        0.24
  5       1873      153.40      156.13       12.06       -2.73       -0.08
  6       1973      160.40      177.74       11.09      -17.34       -0.52
  7       2088      167.90      201.15       10.10      -33.25       -0.98
  8       2208      172.10      194.53        8.46      -22.43       -0.65
  9       2271      183.30      210.07        8.32      -26.77       -0.78
 10       2366      197.50      213.63        8.67      -16.13       -0.47
 11       2423      204.00      184.72       11.55       19.28        0.58
 12       2416      214.50      184.04       11.48       30.46        0.91
 13       2485      228.40      257.85        8.63      -29.45       -0.86
 14       2608      249.30      298.03        9.58      -48.73       -1.43
 15       2744      262.90      273.28       11.15      -10.38       -0.31
 16       2729      274.40      212.39       15.73       62.01        1.96
 17       2695      287.60      263.02       10.77       24.58        0.73
 18       2827      306.40      318.67       11.02      -12.27       -0.37
 19       2959      331.30      353.08       12.39      -21.78       -0.66
 20       3115      358.50      364.21       16.42       -5.71       -0.18
 21       3192      382.90      471.41       14.63      -88.51       -2.75R

 22       3187      408.90      438.54       15.66      -29.64       -0.94
 23       3249      436.50      425.55       25.21       10.95        0.44 X
 24       3166      474.50      434.80       15.60       39.70        1.25
 25       3279      521.20      522.45       15.03       -1.25       -0.04
 26       3501      552.10      588.93       10.62      -36.83       -1.09
 27       3619      620.10      638.88       11.81      -18.78       -0.56
 28       3718      724.70      673.57       16.93       51.13        1.65
 29       3845      750.40      710.14       16.79       40.26        1.30
 30       4017      787.50      762.98       14.34       24.52        0.76
 31       4118      794.80      802.18       21.93       -7.38       -0.27
 32       4156      825.50      809.66       18.51       15.84        0.53

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 1.17
```

Fig. 16.24 MINITAB output for Eq. 16.30

16.10 Regression Approach to Investigating the Effect of Alternative Business Strategies

Johnson et al. (1989) used multiple regression with dummy variables to investigate the relationship between business strategy and wages within the context of a significant environmental change, deregulation of the airline industry (1978–1984).[7] Their regression results were

$$\begin{aligned}\widehat{\log_e \text{Wages}} = {} & \underset{(62.19)}{4.5848^{***}} + \underset{(1.18)}{.003\,\text{PROFITS}} - \underset{(-.19)}{.000\,\text{DEBT}} \\ & + \underset{(2.38)}{.1348\,\text{PERCENT UNION}^{*}} - \underset{(-.78)}{.1250\,\text{LOAD FACTOR}} + \underset{(1.24)}{.0000\,\text{SALES}} \\ & - \underset{(-1.75)}{.1650\,\text{FUEL COST}} - \underset{(-1.88)}{.9040\,\text{COST}^{*}} - \underset{(-1.84)}{.1271\,\text{FOCUS}^{*}} \\ & - \underset{(-2.28)}{.0952\,\text{STUCK}^{*}}\end{aligned} \tag{16.31}$$

Equation 16.31 is a log-linear model discussed in Sect. 16.6. In this estimated multiple regression, t-values appear in parentheses beneath the coefficients, $R^2 = .18$, $n = 92$. *** means $p < .001$; ** means $p < .01$; * means $p < .05$, based on one-sided test.

In this equation, cost, focus, and stuck are business strategic variables as defined in Table 16.11. Table 16.11 presents four alternative business strategies. They are (1) the cost leadership strategy (cost), to maintain the lowest position in the industry; (2) the product differentiation strategy (Diff.), to create a unique product or industry-wide service through brand image (Coca-Cola is a good example), customer service, technology (Polaroid cameras), or other distinguishing features; (3) the focus business strategy (focus), to cater to a narrow strategic target with the aim of being more effective or efficient than those that are competing on a national basis; and (4) the stuck-in-the-middle (stuck) strategy, where no clearly defined strategic position exists.

Results of Eq. 16.31 indicate that estimated coefficients associated with cost, focus, and stuck are all significant at $p = .05$; therefore, we can conclude that different business strategies did affect wages in the airline industry during the years 1978–1984.

[7] This section is essentially based on Johnson, N.B., et al.: Deregulation, business strategy and wages in the airline industry. Ind. Relations **28**(3), 419–430 (1989)

Table 16.11 Strategic classification and mean wages by regulating period

Regulation			Deregulation		
Airline	*Deflated average wage*	*Strategy*	*Airline*	*Deflated average wage*	*Strategy*
US Air	14,099	Focus	National	12,345	Cost
Delta	12,133	Cost	US Air	11,911	Diff.
Ozark	12,094	Focus	American	11,797	Diff.
Frontier	11,959	Focus	Delta	11,400	Cost[a]
Texas Air	11,827	Focus	TWA	11,397	Diff.
American	11,711	Diff.	United	11,240	Diff.
National	11,643	Cost	Western	11,199	Stuck
Eastern	11,374	Stuck	Northwest	11,085	Cost
Piedmont	11,332	Focus	Braniff	11,019	Stuck
TWA	11,201	Diff.	Pan Am	11,017	Stuck
Western	11,104	Stuck	Ozark	11,014	Focus
United	11,010	Diff.	Pacific SW	10,842	Focus
Continental	10,911	Focus	Frontier	10,808	Focus
Pan Am	10,831	Focus	Eastern	10,785	Stuck
Northwest	10,722	Cost	Texas Air	10,423	Cost
Braniff	10,696	Focus	Republic	10,098	Focus
			Continental	9,799	Stuck/Cost[b]
			Piedmont	9,706	Focus
			Southwest	8,902	Focus
			People	4,105	Cost

Source: Based on deregulation, business strategy and wages in the airline industry. Ind. Relations **28**(3), 419–430, reprinted with permission from Basil Blackwell Ltd.
[a]Delta was coded as a Differentiator in alternative regressions
[b]Continental was coded as Stuck in 1978–1982 and as Cost in 1983–1984

16.11 Summary

In this chapter, we extended the concepts and issues of simple and multiple regression that were discussed in Chaps. 13, 14, and 15. Specifically, we investigated other topics in regression analysis, such as multicollinearity, heteroscedasticity, autocorrelation, and misspecification. We also examined nonlinear regression, regression with lagged dependent variables, dummy variables, and interaction variables. Related economics and business examples were used to demonstrate how the new models and techniques presented in this chapter can be used to analyze data.

Questions and Problems

1. Consider the hypothesis that poverty is a function of race and sex. Sample data on the subject are collected and coded using the three dummy variables P, R, and S. The P dummy represents poverty ($P = 1$ for poverty), the R dummy represents race ($R = 1$ for black), and the S dummy represents sex ($S = 1$ for female). Suppose the following multiple regression equation is estimated:

$$\hat{P} = .05 + .23R + .45S$$

(a) Interpret the coefficients of R and S.
(b) Is there any problem associated with this interpretation of the equation? In other words, is there a violation of the assumptions of linear probability models?

2. The relationship between drug abuse and crime has been described by the regression

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + a_3x_3$$

where

x_1 = per-gram retail price of heroin
x_2 = average temperature
x_3 = time trend
y = crime

Say the results for crime are $a_0 = 51.66$, $a_1 = 1.45$ (2.89), $a_2 = .04$ (.22), and $a_3 = .05$ (.07). The values in parentheses are the t-values. $R^2 = .523$.

(a) Interpret the multiple regression equation.
(b) What problems may be associated with the interpretation of this equation? Explain.

3. Protski, Inc., an audit firm, wants to develop a multiple regression model that can explain the value of a house Y, measured in thousands of dollars, by the age of the house X_1, its square footage X_2, the number of bathrooms X_3, the absence (0) or presence (1) of an attached garage D_1, and the absence (0) or presence (1) of a view D_2. A random sample of 20 houses is used to gather observations. Here are the results (standard errors are in parentheses):

$$\begin{aligned}\hat{Y} = 63.53 &- .5827X_1 + .00956X_2 + .81X_3 \\ (38.12)\quad &(.4907)\quad (.01967)\,(11.75) \\ &-4.98D_1 + 13.07D_2 \\ &(19.01)\qquad (17.69)\end{aligned}$$

The error sum of squares is 7,892; the total sum of squares is 9,665.

(a) Comment on the significance of the regression coefficients.
(b) Comment on the overall significance of this regression.

4. (a) Define autocorrelation. State which assumptions of the regression model are violated when autocorrelation exists.
(b) What is the difference between positive and negative autocorrelation?
(c) Describe a technique used to detect autocorrelation.

5. A firm with a nationwide system of bus facilities wants to develop a regression model that can explain its profit Y, measured in thousands of dollars per year, by its annual sales of bus repair and maintenance services (X_1), its annual sales of bus equipment (X_2), and its annual sales of bus advertising panels (X_3). A random sample of 12 of its facilities yields these results (standard errors are in parentheses):

$$\hat{Y} = -2.29 - .0279X_1 + .0885X_2 + 3.753X_3$$
$$\quad (13.65) \quad (.1439) \quad (.0161) \quad (2.402)$$

The error sum of squares is 879.5; the total sum of squares is 5,981.3.

(a) Comment on the significance of the regression coefficients.
(b) Comment on the overall significance of this regression.
(c) Do you see evidence of a possible violation of crucial assumptions?

6. Dividends per share (DPS), price per share (PPS), and retained earnings (RE) for the 30 Dow Jones industrials for 1984 give us the following multiple regression model:

$$\hat{\text{PPS}}_i = 11.336 + 12.434\text{DPS}_i + 3.0875\text{RE}_i$$
$$\quad (2.33) \quad (4.41) \quad (2.39)$$
$$R^2 = .70 \quad F = 31.44$$

Correlation matrix			
Variables	PPS_i, Y	DPS_i, X_1	RE_i, X_2
PPS_i,Y	1	.79761	.69761
DPS_i, X_1	—	1	.62539
RE_i, X_2	—	—	1

(a) Interpret the multiple regression equation.
(b) Interpret the correlation matrix.

7. From Table 16.1, we can define the empirical relationship among PPS_i, DPS_i, and RE_i as

$$\hat{\text{PPS}}_i = 11.336 + 12.434\text{DPS}_i + 3.0875\text{RE}_i$$
$$\quad (\alpha_1) \quad (\beta_1) \quad (\beta_2)$$

We also have
Cov(PPS,DPS) = 20.174
Cov(RE,DPS) = 1.6529
Var(DPS) = 1.2120

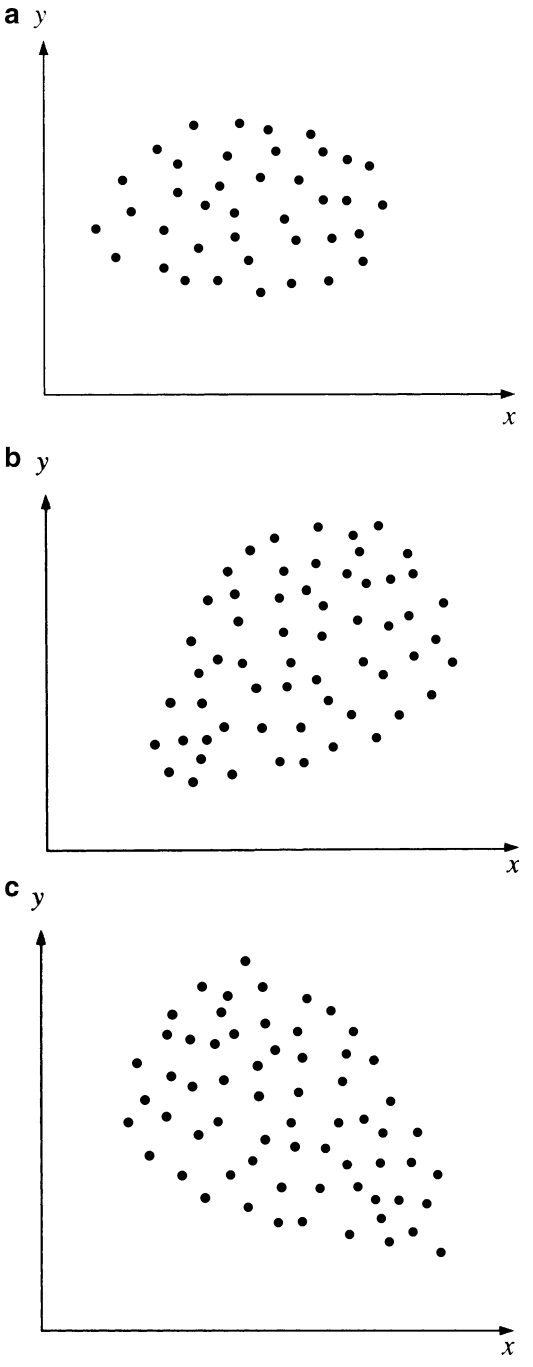
$$\overline{\text{PPS}} = 40.958 \quad \overline{\text{DPS}} = 1.8862 \quad \overline{\text{RE}} = 1.9930$$

(a) Calculate the coefficients of the new equation $\text{PPS}_i = \alpha'_1 + \beta'_1 \text{DPS}_i$.
(b) Regress RE_i on DPS_i and obtain the equation $\text{RE}_i = b_0 + b_1 \text{DPS}_i$. Calculate b_0 and b_1.
(c) Relate β'_1 to β_1 b_1 and β_2, and estimate the specification bias associated with β'_1.

8. What is multicollinearity? What problems does it cause? How can we detect multicollinearity? When we detect multicollinearity, what should we do?
9. What is autocorrelation? What problems does autocorrelation cause? How can we detect autocorrelation?
10. What is heteroscedasticity? What problems does it cause? How can we detect heteroscedasticity?
11. What is specification bias? What problems does specification bias lead to? How can we avoid specification bias?
12. What is a nonlinear regression model? Why do we sometimes choose to estimate a nonlinear model?
13. What is a lagged dependent variable? Why do we use lagged dependent variables in a regression?
14. What is a dummy variable? What does the coefficient on the dummy variable measure? Give some examples drawn from economics, finance, and accounting of times when we would want to use a dummy variable in a regression.
15. Suppose we are interested in measuring the differences in earnings among whites, blacks, Hispanics, and Asians. How many dummy variables should we use in our regression?
16. What are interaction variables? When would we choose to use interaction variables? What does the coefficient of the interaction variable tell us?
17. Suppose you have a sample of 40 observations and 3 explanatory variables and you want to test for autocorrelation. What can you say about autocorrelation if you have the following Durbin–Watson statistics?

(a) $d = 1.30$	(b) $d = 1.00$	(c) $d = 2.25$
(d) $d = 1.95$	(e) $d = 3.55$	

18. When we use a lagged dependent variable in our regression, R^2 is generally much higher than when such a variable is not included. Can you think of any reasons why?
19. Suppose you are interested in how stock returns differ in different months of the year. You decide to use dummy variables to examine this difference. If you choose to use 12 dummy variables, what problem will you encounter? What is the solution to this problem?
20. Look at the following scatter diagrams and explain whether heteroscedasticity appears to be a problem in either of them.


a
y
x
b
y
x
c
y
x

21. When heteroscedasticity is detected, we sometimes use a weighted regression in which the dependent and independent variables are weighted by the variances of their error terms. Thus, the estimated regression becomes $y_i/s_e = \beta_1 x_1/s_e + \beta_2 x_2/s_e + e_i/s_e$, where s_e is the standard error of residuals. Explain intuitively why this may produce better regression results.
22. What assumptions concerning the slope coefficient β must we make when we use dummy variables in a regression?
23. You are interested in the relationship between y and three possible explanatory variables x_1, x_2, and x_3. You are given the following correlation matrix:

	y	x_1	x_2	x_3
y	1.00	.85	.86	.99
x_1		1.00	.32	.85
x_2			1.00	.50
x_3				1.00

Given this information, do you think multicollinearity will be a problem? If so, between which variables?

24. Suppose you have been hired by a lawyer who is interested in showing that a company discriminates against women in the wages it pays. You estimate the regression

$$\hat{\text{WAGE}}_i = 20,000 + 5,000\text{EXPER}_i + 200\text{EDUC}_i - 3,000\text{SEX}_i$$

where

WAGE_i	=	wage for person i
EXPER_i	=	years of experience for person i
EDUC_i	=	years of education for person i
SEX_i	=	dummy variable (1 for female, 0 for male)

(a) Interpret the coefficients for experience and education.
(b) Interpret the coefficient for sex. Does discrimination exist?

25. Suppose you also calculated the standard errors for the coefficients for experience, education, and sex as 2,000, 85, and 2,500, respectively. How would your answer to part (b) of question 24 change?
26. In order to forecast the value of a variable, we sometimes use a nonlinear trend regression such as

$$\hat{y}_t = \alpha + \beta_1 t + \beta_2 t^2 + e_t \tag{A}$$

where t = time. Briefly explain why this model may be better than a model such as

$$\hat{y}_t = \alpha + \beta_1 t + e_t \tag{B}$$

27. Suppose you are given the following data for Abbott Laboratories sales.

Year	Sales	Year	Sales
1968	351.0	1978	1467.6
1969	403.9	1978	1683.2
1970	457.5	1980	2038.2
1971	458.1	1981	2342.5
1972	521.8	1982	2602.4
1973	620.4	1983	2927.9
1974	765.4	1984	3104.0
1975	940.6	1985	3360.3
1976	1084.8	1986	3870.7
1977	1244.9	1987	4387.9

Use MINITAB to do the following:

(a) Use model A from question 26 to estimate the relationship between sales and time (t).
(b) Use model B from question 26 to estimate the relationship among sales, time (t), and the square of time (t^2).

28. Use MINITAB and the data given in question 27 and the equations given in question 26.

(a) Draw a graph showing the actual amount of sales and the estimate of sales based on equation A.
(b) Draw a graph showing the actual amount of sales and the estimate of sales based on equation B.
(c) Referring to the graphs you drew in parts (a) and (b), compare the two models used for forecasting. Which model does a better job?

29. Redo question 28, this time using only data from 1978 to 1987. Does your answer to part (c) change? If so, account for this result.
30. The following are error terms from a regression, where $n = 23$ and $k = 2$.

Year	e	Year	e
1970	1.2	1982	−.50
1971	−.3	1983	−.20
1972	2.4	1984	1.10
1973	−1.0	1985	2.10
1974	.4	1986	−1.50
1975	−.5	1987	2.20
1976	−.4	1988	.50
1977	2.3	1989	−3.10
1978	−2.7	1990	4.20
1979	.1	1991	−1.10
1980	−3.00	1992	1.80
1981	2.40		

Compute the Durbin–Watson d statistic. Does autocorrelation appear to be a problem?

31. A biologist is interested in the effect of temperature and humidity on cell growth. She collects the following information from 8 samples:

Sample	Temperature (°)	Humidity (%)	Number of cells
1	50	20	100
2	55	30	125
3	60	40	175
4	60	50	200
5	70	45	218
6	75	70	235
7	80	65	220
8	85	80	250

Use the MINITAB program to estimate the relationship among number of cells, temperature, and humidity. Use an interaction variable to estimate the interaction effect of temperature and humidity on cell growth. Interpret your results.

32. Use a t test to test the significance of the coefficient on the interaction variable in question 31. Construct a 90 % confidence interval for this coefficient.
33. A popular belief in some financial circles is that most of the movement of the stock market takes place in January. Suppose you are interested in testing this "January effect" on General Motors stock. Explain how you could do this by using a dummy variable.
34. A college admissions officer is interested in knowing whether there is a difference between males' and females' math SAT scores. She collects the following information on the math SAT score, high school grade point average, and sex of 6 students:

Student	Y Math SAT score	X High school GPA	Sex
1	620	3.10	M
2	525	3.85	F
3	650	3.25	M
4	550	3.89	F
5	700	3.60	M
6	675	4.00	F

Use a dummy variable to test whether there is a difference between the math SAT scores of males and females. Be sure to interpret the results.

35. The batting instructor of the Toronto Blue Jays would like to see whether playing ball in the winter (winter ball) has any effect on a player's season batting average. He collects the following information on 6 players:

Player	Y Season batting average	X Hours of batting practice	Played winter ball
1	.300	12	yes
2	.275	11	no
3	.250	8	no
4	.325	20	yes
5	.265	8	yes
6	.350	25	yes

Use a dummy variable to test whether playing winter ball improves a player's regular-season batting average.

36. Suppose you estimate a multiple regression and find the t statistics on the coefficients to be insignificant, whereas the F-statistic indicates that the coefficients are jointly significant. What problem have you probably encountered?
37. Suppose a labor economist is interested in seeing whether there is a difference in earnings and education between people in the northeast and people in other parts of the country. He estimates the following regression:

$$\hat{\text{EARN}}_i = 18,500 + 2,325\text{EDUC}_i + 1,725\text{DUM}_i$$

where DUM_i is a dummy variable equal to 1 if the person is from the northeast and equal to 0 zero if the person is from anywhere else.
Interpret the foregoing regression. Do people from the northeast earn more than people from other parts of the country?

38. A financial analyst is interested in the relationship between dividend per share (DPS) and earnings per share (EPS). He collects information on these two variables and estimates the regression

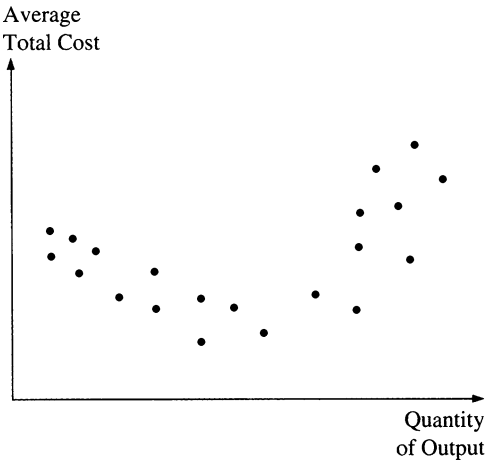
$$\text{DPS}_i = \alpha + \beta\text{EPS}_i + e_i$$

From this regression, he computes the error from the regression for each company.

Company	EPS	e
1	1.00	.05
2	2.10	.35
3	1.20	.10
4	3.00	.72
5	1.25	.25
6	5.20	1.21
7	8.00	2.12

Does heteroscedasticity appear to be a problem in this regression? (Hint: First use MINITAB to plot e_t against EPS.)

39. You are interested in examining the relationship between consumption and income (the consumption function) during two different periods: the period before the Vietnam War and the period during and after the Vietnam War. Explain how you would do this.
40. The following scatter diagram shows the relationship between average total cost and quantity of output. Suggest a multiple regression model to describe this relationship.



41. You are given the following error terms, which are the result of a regression of sales versus advertising expenditures for the Huessy Corporation over a 20-year period:

Year	e_t	Year	e_t
1	.075	11	.065
2	.080	12	.180
3	−.100	13	−.120
4	−.070	14	−.070
5	.500	15	.050
6	−.230	16	−.230
7	−.007	17	−.107
8	.088	18	.288
9	−.101	19	−.131
10	−.007	20	−.007

(a) Compute the Durbin–Watson statistic.
(b) Does autocorrelation exist?

Year	Chrys.	Ford	GM	Indus.
Debt/equity ratio				
69	1.23	.76	.45	.74
70	1.23	.81	.44	.79
71	1.20	.89	.69	.91
72	1.21	.95	.56	.89
73	1.24	1.02	.62	.92
74	1.53	1.27	.63	1.09
75	1.60	1.21	.66	1.10
76	1.51	1.22	.70	1.08

(continued)

(continued)

Year	Chrys.	Ford	GM	Indus.
77	1.62	1.28	.69	1.12
78	1.39	1.28	.74	1.14
79	2.65	1.26	.68	1.23
80	13.41	1.84	.94	2.42
81	7.04	2.13	1.20	2.07
82	5.32	2.61	1.26	7.03
83	3.96	2.16	1.20	2.14
84	1.74	1.79	1.15	1.66
85	1.99	1.58	1.16	2.57
86	1.71	1.55	1.37	2.07
87	2.07	1.43	1.63	1.90
88	5.41	5.66	3.60	4.61
Return on assets				
69	.02	.06	.11	.08
70	.00	.05	.04	.04
71	.02	.06	.11	.07
72	.04	.07	.12	.09
73	.04	.07	.12	.09
74	−.01	.03	.05	.03
75	−.03	.02	.06	.03
76	.05	.06	.12	.08
77	.02	.09	.12	.09
78	−.03	.07	.11	.07
79	−.17	.05	.09	.05
80	−.26	−.06	−.02	−.05
81	−.08	−.05	.01	−.02
82	−.02	−.03	.02	−.01
83	.04	.08	.08	.07
84	.16	.11	.09	.10
85	.13	.08	.06	.07
86	.10	.09	.04	.06
87	.06	.10	.04	.06
88	.02	.04	.03	.03

Use MINITAB and the following information to answer questions 42–54. To find out whether there is a relationship between the amount of financial leverage a firm uses and the return on the firm's assets, you collect information on the debt/equity ratio and return on assets for the "big three" automakers and the average for the auto industry.

42. Estimate the regression of Chrysler's return on assets against its debt/equity ratio. Compute the Durbin–Watson statistic. Does autocorrelation exist?
43. Redo question 42 using the data for Ford.
44. Redo question 42 using the data for GM.
45. Redo question 42 using the data for the auto industry.

46. Using Chrysler's data, compute r_1, the correlation coefficient between e_t and e_{t-1}, for the error terms computed in Chrysler's regression model.
47. Redo question 46 using the data for Ford.
48. Redo question 46 using the data for GM.
49. Redo question 46 using the data for the auto industry.
50. Compare your computations from questions 46–49 with the Durbin–Watson statistics you calculated in questions 42–45.
51. Suppose you believe that in addition to there being a relationship between the return on assets and the debt/equity ratio, the return on assets in the previous period may also play an important part in determining the return on assets in the current period. You estimate the equation:

$$\text{ROA}_t = a + b\text{DE}_t + \text{ROA}_{t-1} + e_i$$

    (a) Use the data for Chrysler to estimate the foregoing equation.
    (b) Compare your results to the simple regressions you computed in question 42.

52. Redo question 51 using the data for Ford.
53. Redo question 51 using the data for GM.
54. Redo question 51 using the data for the auto industry.
55. Suppose you have a sample of 50 observations and 4 explanatory variables and you want to test for autocorrelation. What can you say about autocorrelation if you have the following Durbin–Watson statistics?

(a) $d = 1.90$	(b) $d = .90$	(c) $d = 2.55$
(d) $d = 1.75$	(e) $d = 3.45$	

56. A financial analyst is interested in the relationship between earnings per share (EPS) and sales. She collects information on these two variables. Then she estimates the regression

$$\text{EPS}_i = \alpha + \beta\text{SALES}_i + e_i$$

and, from it, computes the error from the regression for each company.

Company	Sales	e
1	1,200	400.05
2	2,210	500.35
3	3,201	− 50.10
4	3,400	−200.72
5	4,525	300.25
6	5,320	−100.21
7	6,001	−87.25

Does heteroscedasticity appear to be a problem in this regression?

57. Suppose we have the following two versions of the market model, which shows the relationship between the rate of return on a stock and the rate of return on some market index:
Standard market model: $R_{i,t} = \alpha + \beta R_{m,t} + e_{i,t}$
Quadratic market model: $R_{i,t} = \alpha + \beta R_{m,t} + \gamma R_{m,t}^2 + e_{i,t}$

    (a) Suppose the quadratic market model is the correct form to estimate but we estimate the standard market model instead. What is the effect on our parameter estimates?
    (b) Suppose the standard market model is the correct form to estimate but we estimate the quadratic market model instead. What is the effect on our parameter estimates?

58. An economist estimates the equation:

$$\hat{C}_t = 1,000 + .75Y_t + .10C_{t-1}$$

where

C_t = consumption in time period t
Y_t = income in time period t
C_{t-1} = consumption in time period $t - 1$

Interpret the coefficients of the regression model.

59. A farmer is interested in measuring the relationship among the number of bushels of corn grown, the amount of rainfall, and the amount of fertilizer used. Give two different equations that he could use to find this relationship.
60. Suppose the farmer in question 59 collects the following data:

Bushels of corn	Rainfall (inches)	Fertilizer (pounds)
1,000	4	10
1,211	5	9
1,600	7	12
900	2	4
2,000	9	15

    (a) Estimate the models you suggested in question 59.
    (b) Compare the results of the two models. Which model do you believe is better? Explain.

61. A marketing manager believes that advertising expenditures are effective in increasing sales only to a certain extent. He discovers that when the advertising expenditures exceed a certain level, sales respond accordingly. Propose two mathematical models that can be used to describe this type of data.
62. An economist wants to study the factors that determine the hourly wage rate. He comes up with the regression

$$\hat{y} = a + bx_1 + cx_2 + dx_3 + e$$

where

y = hourly wage rate
x_1 = age of the employee
x_2 = years of experience
x_3 = years of schooling

What problem might the economist encounter in estimating this model?

63. An economist conducts research on the relationship between household spending and income. She collects data on 60 households' spending and income during 1986. After using regression analysis, she obtains the following results:

$$\hat{\text{CONSUMPTION}} = 4.906 + .756 \times \text{INCOME}$$
$$R^2 = .58 \quad \text{DW} = .32$$

The researcher claims that there is a serial correlation problem because of the low DW statistic. Do you believe this conclusion is correct?

64. Which of the following models is nonlinear in parameters where y and x are variables and a, b, and c are parameters?

(a) $y = ax_1^b x_2^c$	(c) $y = a + bx + cx^2$
(b) $y = a + bx^c$	(d) $y = a + b(1/x)$

Use the following information to answer questions 65–70. What determines the voting behavior in a presidential election? An economist believes that people "vote their pockets." He argues that economic condition is the greatest concern of voters. Therefore, the percentage of votes the incumbent gets depends on macroeconomic variables such as the inflation rate (IR), the unemployment rate (UR), and the growth rate of disposable income (DI). The following table contains these data:

y	DI	UR	IR
45	1.2	8.3	3.5
48	1.8	7.4	3.8
49.5	2.0	7.1	3.9
48.8	1.9	6.5	4.2
50.4	2.2	6.2	4.6
51.3	2.4	5.9	4.9
52.4	2.7	5.7	5.0
47.6	1.9	7.0	3.4
54.1	3.0	5.1	5.1
50.0	2.3	6.1	3.4

In the regression, we are interested in what percentage of votes the incumbent receives. We decide to use the following model:

$$y = a + b_1\text{DI} + b_2\text{UR} + b_3\text{IR} + e$$

where

y = percentage of votes the incumbent receives
DI = rate of increase in disposable income
UR = unemployment rate
IR = inflation rate

65. What are the hypothesized signs of b_1, b_2, and b_3?
66. Estimate the model using regression analysis. Conduct a test to see whether the coefficients of UR and IR are significant at the 5 % level.
67. Use an F test to test the hypothesis that $b_2 = b_3 = 0$ at the 5 % level.
68. When Jimmy Carter was a candidate for the presidency, he coined a new term, the Misery Index, where Misery Index = UR + IR. Run a regression with only DI and the Misery Index. Conduct a test at the 5 % level to see whether the coefficient on the misery index is significant in explaining y.
69. Are you convinced that DI is the only significant variable that affects y in our model? Before you run the regression, do you expect to obtain a low or a high R^2?
70. A political scientist wants to add an important variable to the equation associated with question 64 to catch the effect of war on voting behavior. He argues that as a result of patriotism, the incumbent receives a higher percentage of votes during a war than during peacetime. Suggest a way to specify the new model in order to catch the patriotism effect.
71. A financial analyst is interested in the relationship between earnings per share (EPS) and sales for Addison Company. He collects information on these two variables for a 10-year period. He estimates the regression

$$\text{EPS}_i = \alpha + \beta\text{SALES}_i + e_i$$

and from it computes the error from the regression for each company.

Year	e_t	Year	e_t
1	130.05	6	−10.11
2	−540.35	7	83.35
3	150.10	8	−90.30
4	− 240.32	9	34.04
5	100.24	10	−127.20

Does autocorrelation appear to be a problem in this regression?

72. Based on the correlations among the returns for T-Bill, Chrysler, Ford, GM, and NYSE Index in the following output, is there a problem about multicollinearity?

	T-Bill	Chrysler	Ford	GM
Chrysler	0.066			
Ford	−0.098	0.814		
GM	−0.089	0.767	0.835	
NYSE Index	0.084	0.831	0.842	0.806

73. The Durbin–Watson statistic computed from fitting a regression on Ford using NYSI Index and Chrysler as predictors is 1.9904. Is there a problem of autocorrelation?
74. (Problem *73* continued.) From the following result, is there a "January effect" on returns of Ford *even if NYSI Index and Chrysler are included as predictors*?

Predictor	Coef	SE Coef
Constant	0.009900	0.008017
NYSE index	0.7207	0.2364
Chrysler	0.3027	0.1107
Jan	0.07460	0.02872

75. (Problem *74* continued.) Write down the estimated regression equations for January and non-January, respectively.
76. (Problem *75* continued.) Since the January effect is significant, do we need to consider the interaction between the January effect and the returns of NYSE Index and Chrysler?
77. Add an interaction term (GNP) (Dummy) to Eq. 16.26. The MINITAB outputs are presented as follows. Please compare these results with those presented in Table 16.10.

```
MTB >          BRIEF 3
MTB >          REGRESS 'M3' 4 'GNP' 'PRIMERt' 'DUMMY' 'GNPDUMMY';
SUBC>          DW.
```

Regression Analysis

The regression equation is

M3 = − 693 + 0.383 GNP − 6.47 PRIMERt + 565 DUMMY − 0.218 GNPDUMMY

Predictor	Coef	StDev	T	P
Constant	−693.0	122.1	−5.67	0.000
GNP	0.38348	0.02719	14.11	0.000
PRIMERt	−6.468	2.924	−2.21	0.036
DUMMY	565.3	128.7	4.39	0.000
GNPDUMMY	−0.21835	0.03634	−6.01	0.000

S = 26.84 R-Sq = 98.7% R-Sq(adj) = 98.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	1507410	376853	523.03	0.000
Error	27	19454	721		
Total	31	1526864			

Source	DF	Seq SS
GNP	1	1391385
PRIMERt	1	19320
DUMMY	1	70686
GNPDUMMY	1	26019

Obs	GNP	M3	Fit	StDev Fit	Residual	St Resid
1	1629	140.00	112.34	11.36	27.56	1.14
2	1665	140.70	116.12	11.03	24.58	1.00
3	1709	145.20	125.35	10.40	19.85	0.80
4	1799	147.90	140.33	9.41	7.57	0.30
5	1873	153.40	152.53	8.71	0.87	0.03
6	1973	160.40	169.05	7.97	−8.65	−0.34
7	2088	167.90	187.66	7.43	−19.76	−0.77
8	2208	172.10	200.54	6.28	−28.44	−1.09
9	2271	183.30	211.09	6.22	−27.79	−1.06
10	2366	197.50	222.19	6.03	−24.69	−0.94
11	2423	204.00	220.98	7.35	−16.98	−0.66
12	2416	214.50	220.13	7.29	−5.63	−0.22
13	2485	228.40	245.62	7.07	−17.22	−0.67
14	2608	249.30	269.09	9.04	−19.79	−0.78
15	2744	262.90	273.50	8.36	−10.60	−0.42
16	2729	274.40	253.07	12.55	21.33	0.90
17	2695	287.60	266.49	7.91	21.11	0.82
18	2827	306.40	294.83	9.54	11.57	0.46
19	2959	331.30	316.68	11.29	14.62	0.60
20	3115	358.50	328.11	12.34	30.39	1.27
21	3192	382.90	449.32	12.13	−66.42	−2.77R
22	3187	408.90	430.47	12.05	−21.57	−0.90
23	3249	436.50	430.84	18.42	5.66	0.29X
24	3166	474.50	425.03	12.02	49.47	2.06R
25	3279	521.20	494.72	12.96	26.48	1.13
26	3501	552.10	571.89	7.99	−19.79	−0.77
27	3619	620.10	630.52	9.16	−10.42	−0.41
28	3718	724.70	678.91	11.37	45.79	1.88
29	3845	750.40	728.48	11.25	21.92	0.90
30	4017	787.50	787.17	12.00	0.33	0.01
31	4118	794.80	815.80	15.10	−21.00	−0.95
32	4156	825.50	835.97	15.06	−10.47	−0.47

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence
Durbin-Watson statistic = 1.23

78. Add an interaction term (GNP) (Dummy) to Eq. 16.29. The MINITAB output is presented as follows. Please compare these results with those presented in Fig. 16.22.

```
MTB >    BRIEF 3
MTB >    REGRESS 'M3' 5 'GNP' 'PRIMERt' 'DUMMY' 'GNPPRIME' 'GNPDUMMY'
SUBC>    DW
```

Regression Analysis

The regression equation is

M3 = − 754 + 0.403 GNP + 0.1 PRIMERt + 598 DUMMY − 0.00210 GNPPRIME
− 0.230 GNPDUMMY

Predictor	Coef	StDev	T	P
Constant	−753.6	228.8	−3.29	0.003
GNP	0.40312	0.06807	5.92	0.000
PRIMERt	0.09	20.98	0.00	0.997
DUMMY	598.5	167.9	3.56	0.001
GNPPRIME	−0.002097	0.006644	−0.32	0.755
GNPDUMMY	−0.23016	0.05259	−4.38	0.000

S = 27.30 R-Sq = 98.7% R-Sq(adj) = 98.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	1507485	301497	404.49	0.000
Error	26	19380	745		
Total	31	1526864			

Source	DF	Seq SS
GNP	1	1391385
PRIMERt	1	19320
DUMMY	1	70686
GNPPRIME	1	11815
GNPDUMMY	1	14278

Obs	GNP	M3	Fit	StDev Fit	Residual	St Resid
1	1629	140.00	111.70	11.73	28.30	1.15
2	1665	140.70	116.46	11.27	24.24	0.97
3	1709	145.20	124.65	10.82	20.55	0.82
4	1799	147.90	139.48	9.95	8.42	0.33
5	1873	153.40	151.56	9.38	1.84	0.07
6	1973	160.40	167.91	8.87	−7.51	−0.29
7	2088	167.90	186.43	8.51	−18.53	−0.71
8	2208	172.10	201.20	6.72	−29.10	−1.10
9	2271	183.30	211.47	6.44	−28.17	−1.06
10	2366	197.50	223.29	7.06	−25.79	−0.98

(continued)

(continued)

11	2423	204.00	224.21	12.69	−20.21	−0.84
12	2416	214.50	223.35	12.63	−8.85	−0.37
13	2485	228.40	245.30	7.26	−16.90	−0.64
14	2608	249.30	267.74	10.14	−18.44	−0.73
15	2744	262.90	273.95	8.62	−11.05	−0.43
16	2729	274.40	255.97	15.72	18.43	0.83
17	2695	287.60	267.23	8.38	20.37	0.78
18	2827	306.40	293.79	10.25	12.61	0.50
19	2959	331.30	314.77	12.98	16.53	0.69
20	3115	358.50	325.24	15.50	33.26	1.48
21	3192	382.90	449.58	12.36	−66.68	−2.74R
22	3187	408.90	430.44	12.25	−21.54	−0.88
23	3249	436.50	429.12	19.51	7.38	0.39
24	3166	474.50	425.29	12.26	49.21	2.02R
25	3279	521.20	494.99	13.21	26.21	1.10
26	3501	552.10	570.51	9.23	−18.41	−0.72
27	3619	620.10	630.66	9.33	−10.56	−0.41
28	3718	724.70	680.92	13.20	43.78	1.83
29	3845	750.40	730.93	13.83	19.47	0.83
30	4017	787.50	787.98	12.47	−0.48	−0.02
31	4118	794.80	813.39	17.15	−18.59	−0.88
32	4156	825.50	835.30	15.47	−9.80	−0.44

R denotes an observation with a large standardized residual
Durbin-Watson statistic = 1.22

Project IV: Project for Regression and Correlation Analyses

Use PPS, EPS, and DPS for the Dow Jones' 30 firms presented in Tables IV.1A, IV.1B, IV.1C, and IV.1D to do the following:

1. Calculate the annual rate of return for all 30 firms, and calculate all statistics presented in Table 9.1.
2. Use both the annual rates of return for the 30 firms obtained in (1) and the annual market rates of return presented in Table 2.4 to estimate market models for all 30 firms.
3. (a) Estimate the following regression in accordance with estimates obtained in (1) and (2) where $\bar{R}_i$ and β_i are average rates of return and the estimated beta coefficient respectively: $\bar{R}_i = \alpha + b\beta_i + e_i$.
   (b) Use Eq. 14.36 to adjust for the estimate of b.

(continued)

Project IV: (continued)

4. Use the data presented in Tables IV.1A, IV.1B, IV.1C, and IV.1D to estimate Eq. 16.13 for all 30 firms.
5. Use the data presented in Tables IV.1A, IV.1B, IV.1C, and IV.1D to estimate Eq. 16.13a for all 30 firms

Download monthly adjusted close price data of JNJ and S&P 500 index from Yahoo Finance during the period from January 2005 to current month to do the following:

6. Calculate the monthly rates of return for JNJ and S&P 500 index, and calculate all statistics presented in Table 9.1.
7. Use both monthly rates of return for JNJ and S&P 500 to estimate market model.

Table IV.1A PPS, EPS, and DPS for Dow Jones 30 industrial firms (1990–2009)

	Alcoa Inc			American Express Co			Verizon Communications Inc			Boeing Co		
Year	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS
1990	57.625	3.40	3.050	20.625	0.69	0.920	53.625	3.38	2.320	45.375	4.01	0.950
1991	64.375	0.71	1.780	20.500	1.59	0.940	48.250	3.41	2.480	47.750	4.56	1.000
1992	71.625	0.24	1.600	24.875	0.83	1.000	51.250	3.23	2.580	40.125	4.57	1.000
1993	69.375	0.03	1.600	30.875	3.17	1.000	59.250	3.39	2.660	43.250	3.66	1.000
1994	86.625	4.96	1.600	29.500	2.68	0.950	49.750	3.21	2.740	47.000	2.51	1.000
1995	52.875	4.43	0.900	41.375	3.11	0.900	66.875	4.25	2.790	78.375	1.15	1.000
1996	63.750	2.94	1.330	56.500	3.90	0.900	64.750	3.96	2.860	106.500	3.19	1.090
1997	70.375	4.66	0.975	89.250	4.29	0.900	91.000	3.16	2.970	48.937	−0.18	0.560
1998	74.563	4.87	1.500	102.500	4.71	0.900	54.000	1.90	1.540	32.625	1.16	0.560
1999	83.000	2.87	0.805	166.250	5.54	0.900	61.563	2.72	1.540	41.438	2.52	0.560
2000	33.500	1.83	0.500	54.938	2.12	0.315	50.125	3.98	1.540	66.000	2.48	0.560
2001	35.550	1.06	0.600	35.690	0.99	0.320	47.460	0.22	1.540	38.780	3.46	0.680
2002	22.780	0.59	0.600	35.350	2.02	0.320	38.750	1.67	1.540	32.990	2.90	0.680
2003	38.000	1.21	0.600	48.230	2.34	0.360	35.080	1.27	1.540	42.140	0.90	0.680
2004	31.420	1.61	0.600	56.370	2.79	0.420	40.510	2.62	1.540	51.770	2.27	0.770
2005	29.570	1.41	0.600	51.460	2.61	0.480	30.120	2.67	1.600	70.240	3.26	1.000
2006	30.010	2.49	0.600	60.670	3.08	0.540	37.240	1.88	1.620	88.840	2.88	1.200
2007	36.550	2.98	0.680	52.020	3.45	0.600	43.690	1.90	1.645	87.460	5.36	1.400
2008	11.260	0.28	0.680	18.550	2.49	0.720	33.900	2.26	1.750	42.670	3.68	1.600
2009	16.120	−1.06	0.230	40.520	1.55	0.720	33.130	1.29	1.855	54.130	1.89	1.680

Source: EPS, DPS, and PPS for Dow Jones 30 are from Standard & Poor's Compustat, Wharton Research Data Services (WRDS)

Table IV.1B PPS, EPS, and DPS for Dow Jones 30 industrial firms (1990–2009)

	Caterpillar Inc			JPMorgan Chase & Co			Chevron Corp			Coca-Cola Co		
Year	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS
1990	47.000	2.07	1.200	10.750	2.38	2.720	72.625	6.10	2.950	46.500	2.04	0.800
1991	43.875	−4.00	1.200	21.250	0.11	1.000	69.000	3.69	3.250	80.250	2.43	0.960
1992	53.625	−2.16	0.600	38.625	3.90	1.199	69.500	6.52	3.300	41.875	1.43	0.560
1993	89.000	6.72	0.600	40.125	5.63	1.290	87.125	3.89	3.500	44.625	1.68	0.680
1994	55.125	4.70	0.450	35.875	4.64	1.580	44.625	2.60	1.850	51.500	1.98	0.780
1995	58.750	5.72	1.200	58.750	6.23	1.880	52.375	1.43	1.925	74.250	2.37	0.880
1996	75.250	7.07	1.500	89.375	5.02	2.180	65.000	3.99	2.080	52.625	1.40	0.500
1997	48.500	4.44	0.900	109.500	8.30	2.420	77.000	4.97	2.280	66.687	1.67	0.560
1998	46.000	4.17	1.100	71.000	4.35	1.030	82.938	2.05	2.440	67.000	1.43	0.600
1999	47.063	2.66	1.250	77.688	6.49	1.590	86.625	3.16	2.480	58.250	0.98	0.640
2000	47.313	3.04	1.330	45.438	2.99	1.233	84.438	7.98	2.600	60.938	0.88	0.680
2001	52.250	2.35	1.380	36.350	0.84	1.340	89.610	3.71	2.650	47.150	1.60	0.720
2002	45.720	2.32	1.400	24.000	0.81	1.360	66.480	1.07	2.800	43.840	1.60	0.800
2003	83.020	3.18	1.420	36.730	3.32	1.360	86.390	7.15	2.860	50.750	1.77	0.880
2004	97.510	5.95	1.560	39.010	1.59	1.360	52.510	6.16	1.530	41.640	2.00	1.000
2005	57.770	4.21	0.910	39.690	2.43	1.360	56.770	6.58	1.750	40.310	2.04	1.120
2006	61.330	5.37	1.100	48.300	3.93	1.360	73.530	7.84	2.010	48.250	2.16	1.240
2007	72.560	5.55	1.320	43.650	4.51	1.440	93.330	8.83	2.260	61.370	2.59	1.360
2008	44.670	5.83	1.560	31.530	0.86	1.520	73.970	11.74	2.530	45.270	2.51	1.520
2009	56.990	1.45	1.680	41.670	2.25	0.530	76.990	5.26	2.660	57.000	2.95	1.640

Source: EPS, DPS, and PPS for Dow Jones 30 are from Standard & Poor's Compustat, Wharton Research Data Services (WRDS)

Table IV.1C PPS, EPS, and DPS for Dow Jones 30 industrial firms (1990–2009)

	Disney (Walt) Co			Du Pont (E I) De Nemours			Exxon Mobil Corp			General Electric Co		
Year	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS
1990	90.625	6.00	0.530	36.750	3.40	1.620	51.750	3.96	2.470	57.375	4.85	1.880
1991	114.000	4.78	0.640	46.625	2.08	1.680	60.875	4.45	2.680	76.500	5.10	2.040
1992	36.250	1.52	0.193	47.125	1.43	1.740	61.125	3.82	2.830	85.500	5.02	2.240
1993	37.750	1.23	0.230	48.250	0.83	1.760	63.125	4.21	2.880	104.875	5.18	2.520
1994	38.750	2.04	0.275	56.125	4.00	1.820	60.750	4.07	2.910	51.000	3.46	1.440
1995	57.375	2.60	0.330	69.875	5.61	2.030	80.500	5.18	3.000	72.000	3.90	1.640
1996	63.250	1.96	0.400	94.125	6.47	2.230	98.000	6.02	3.120	98.875	4.40	1.840
1997	80.625	2.86	0.485	60.062	2.12	1.230	61.187	3.41	1.625	73.375	2.50	1.040
1998	25.375	0.91	0.193	53.063	1.45	1.365	73.125	2.64	1.640	102.000	2.84	1.200
1999	26.000	0.63	0.053	65.875	0.19	1.400	80.563	2.28	1.670	154.750	3.27	1.400
2000	38.250	0.58	0.210	48.313	2.21	1.400	86.938	4.60	1.760	47.938	1.29	0.547
2001	18.620	0.11	0.210	42.510	4.17	1.400	39.300	2.20	0.910	40.080	1.42	0.640
2002	15.140	0.61	0.210	42.400	1.84	1.400	34.940	1.62	0.920	24.350	1.52	0.720
2003	20.170	0.65	0.210	45.890	1.00	1.400	41.000	3.16	0.980	30.980	1.56	0.760
2004	22.550	1.14	0.210	49.050	1.78	1.400	51.260	3.91	1.060	36.500	1.62	0.800
2005	24.130	1.27	0.240	42.500	2.08	1.460	56.170	5.76	1.140	35.050	1.76	0.880
2006	30.910	1.68	0.270	48.710	3.41	1.480	76.630	6.68	1.280	37.210	1.99	1.000
2007	34.390	2.33	0.310	44.090	3.25	1.520	93.690	7.36	1.370	37.070	2.21	1.120
2008	30.690	2.34	0.350	25.300	2.21	1.640	79.830	8.78	1.550	16.200	1.79	1.240
2009	27.460	1.78	0.350	33.670	1.93	1.640	68.190	3.99	1.660	15.130	1.03	0.820

Source: EPS, DPS, and PPS for Dow Jones 30 are from Standard & Poor's Compustat, Wharton Research Data Services (WRDS)

Table IV.1D PPS, EPS, and DPS for Dow Jones 30 industrial firms (1990–2009)

	Hewlett-Packard Co			Home Depot Inc			Intel Corp			Intl Business Machines Corp		
Year	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS
1990	26.000	3.06	0.420	35.125	1.42	0.110	38.500	3.20	0.000	113.000	10.51	4.840
1991	50.375	3.02	0.480	43.250	1.35	0.110	49.000	3.92	0.000	89.000	−0.99	4.840
1992	56.875	3.49	0.725	61.625	1.20	0.110	87.000	4.97	0.100	50.375	−12.03	4.840
1993	73.625	4.65	0.900	65.125	1.09	0.110	62.000	5.20	0.200	56.500	−14.02	1.580
1994	97.875	6.14	1.100	39.000	1.01	0.113	63.875	5.24	0.225	73.500	5.02	1.000
1995	92.625	4.63	0.700	46.750	1.32	0.150	56.750	4.03	0.140	91.375	7.23	1.000
1996	44.125	2.46	0.440	46.000	1.54	0.190	130.937	5.81	0.180	151.500	10.24	1.300
1997	61.625	2.95	0.520	49.500	1.94	0.230	70.250	4.25	0.085	104.625	6.18	0.775
1998	60.250	2.85	0.600	60.500	1.59	0.190	118.563	3.64	0.130	184.375	6.75	0.860
1999	74.188	3.08	0.640	60.500	1.10	0.115	82.313	2.20	0.110	107.875	4.25	0.470
2000	46.500	1.80	0.320	56.625	1.03	0.113	30.063	1.57	0.070	85.000	4.58	0.510
2001	16.830	0.32	0.320	48.200	1.11	0.160	31.450	0.19	0.080	120.960	4.45	0.550
2002	15.800	−0.37	0.320	50.090	1.30	0.170	15.570	0.47	0.080	77.500	3.13	0.590
2003	22.310	0.83	0.320	20.900	1.57	0.210	32.050	0.86	0.080	92.680	4.42	0.630
2004	18.660	1.16	0.320	35.470	1.88	0.260	23.390	1.17	0.160	98.580	5.04	0.700
2005	28.040	0.83	0.320	41.260	2.27	0.325	24.960	1.42	0.320	82.200	4.99	0.780
2006	38.740	2.23	0.320	40.550	2.73	0.400	20.250	0.87	0.400	97.150	6.15	1.100
2007	51.680	2.76	0.320	40.740	2.80	0.675	26.660	1.20	0.450	108.100	7.32	1.500
2008	38.280	3.35	0.320	30.640	2.28	0.900	14.660	0.93	0.548	84.160	9.07	1.900
2009	47.460	3.21	0.320	21.530	1.37	0.900	20.400	0.79	0.560	130.900	10.12	2.150

Source: EPS, DPS, and PPS for Dow Jones 30 are from Standard & Poor's Compustat, Wharton Research Data Services (WRDS)

Table IV.1E PPS, EPS, and DPS for Dow Jones 30 industrial firms (1990–2009)

	Johnson & Johnson			McDonald's Corp			Merck & Co			3M Co		
Year	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS
1990	71.750	3.43	1.310	29.125	2.20	0.333	89.875	4.56	1.910	85.750	5.91	2.920
1991	114.500	4.39	1.540	38.000	2.35	0.363	166.500	5.49	2.310	95.250	5.26	3.120
1992	50.500	2.46	0.890	48.750	2.60	0.393	43.375	2.12	0.920	100.625	5.65	3.200
1993	44.875	2.74	1.010	57.000	2.91	0.423	34.375	1.87	1.030	108.750	5.82	3.320
1994	54.750	3.12	1.130	29.250	1.68	0.234	38.125	2.38	1.140	53.375	3.13	1.760
1995	85.500	3.72	1.280	45.125	1.97	0.263	65.625	2.70	1.240	66.375	3.11	1.880
1996	49.750	2.17	0.735	45.375	2.21	0.293	79.625	3.20	1.420	83.000	3.63	1.920
1997	65.875	2.47	0.850	47.750	2.35	0.323	106.000	3.83	1.690	82.062	5.14	2.120
1998	83.875	2.27	0.970	76.813	2.28	0.353	147.500	4.41	1.890	71.125	3.01	2.200
1999	93.250	3.00	1.090	40.313	1.44	0.195	67.188	2.51	1.100	97.875	4.39	2.240
2000	105.063	3.45	1.240	34.000	1.49	0.215	93.625	2.96	1.210	120.500	4.69	2.320
2001	59.100	1.87	0.700	26.470	1.27	0.225	58.800	3.18	1.370	118.210	3.63	2.400
2002	53.710	2.20	0.795	16.080	0.78	0.235	56.610	3.17	1.410	123.300	5.06	2.480
2003	51.660	2.42	0.925	24.830	1.19	0.400	46.200	2.95	1.450	85.030	3.07	1.320
2004	63.420	2.87	1.095	32.060	1.81	0.550	32.140	2.62	1.490	82.070	3.83	1.440
2005	60.100	3.50	1.275	33.720	2.06	0.670	31.810	2.11	1.520	77.500	4.23	1.680
2006	66.020	3.76	1.455	44.330	2.33	1.000	43.600	2.04	1.520	77.930	5.15	1.840
2007	66.700	3.67	1.620	58.910	1.96	1.500	58.110	1.51	1.140	84.320	5.70	1.920
2008	59.830	4.62	1.795	62.190	3.83	1.625	30.400	3.66	1.520	57.540	4.95	2.000
2009	64.410	4.45	1.930	62.440	4.17	2.050	36.540	5.67	1.520	82.670	4.56	2.040

Source: EPS, DPS, and PPS for Dow Jones 30 are from Standard & Poor's Compustat, Wharton Research Data Services (WRDS)

Table IV.1F PPS, EPS, and DPS for Dow Jones 30 industrial firms (1990–2009)

	Bank of America Corp			Pfizer Inc			Procter & Gamble Co			AT&T INC		
Year	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS
1990	22.875	3.40	1.420	80.750	4.77	2.400	87.125	4.49	1.750	56.000	3.67	2.720
1991	40.625	0.76	1.480	84.000	2.13	1.320	77.375	4.92	1.950	64.625	3.85	2.820
1992	51.375	4.60	1.510	72.500	3.25	1.480	46.000	2.62	1.025	74.000	4.34	2.900
1993	49.000	5.00	1.640	69.000	2.05	1.680	52.000	0.25	1.100	41.500	2.39	1.498
1994	45.125	6.12	1.880	77.250	4.19	1.880	53.375	3.09	1.240	40.375	2.74	1.563
1995	69.625	7.13	2.080	63.000	2.47	1.040	71.875	3.71	1.400	57.250	3.10	1.633
1996	97.750	8.00	2.400	83.000	2.99	1.200	90.625	4.29	1.600	51.875	3.46	1.703
1997	60.812	4.27	1.370	74.562	1.76	0.680	141.250	4.87	1.800	73.250	1.62	1.773
1998	60.125	2.97	1.590	125.000	1.54	0.760	91.063	2.74	1.010	53.625	2.08	0.925
1999	50.188	4.56	1.850	32.438	0.85	0.453	89.250	2.75	1.140	48.750	1.93	0.965
2000	45.875	4.56	2.060	46.000	0.60	0.360	57.250	2.61	1.280	47.750	2.35	1.005
2001	62.950	4.26	2.280	39.850	1.25	0.440	63.800	2.15	1.400	39.170	2.16	1.023
2002	69.570	6.08	2.440	30.570	1.49	0.520	89.300	3.26	1.520	27.110	2.24	1.066
2003	80.430	7.27	2.880	35.330	0.22	0.600	89.180	3.90	1.640	26.070	1.80	1.368
2004	46.990	3.76	1.700	26.890	1.51	0.680	54.440	2.46	0.933	25.770	1.50	1.250
2005	46.150	4.10	1.900	23.320	1.10	0.760	52.750	2.83	1.030	24.490	1.42	1.290
2006	53.390	4.66	2.120	25.900	1.52	0.960	55.600	2.79	1.150	35.750	1.89	1.330
2007	41.260	3.35	2.400	22.730	1.19	1.160	61.190	3.22	1.280	41.560	1.95	1.420
2008	14.080	0.56	2.240	17.710	1.19	1.280	60.810	3.86	1.450	28.500	2.17	1.600
2009	15.060	−0.29	0.040	18.190	1.23	0.800	51.100	3.76	1.640	28.030	2.12	1.640

Source: EPS, DPS, and PPS for Dow Jones 30 are from Standard & Poor's Compustat, Wharton Research Data Services (WRDS)

Table IV.1G PPS, EPS, and DPS for Dow Jones 30 industrial firms (1990–2009)

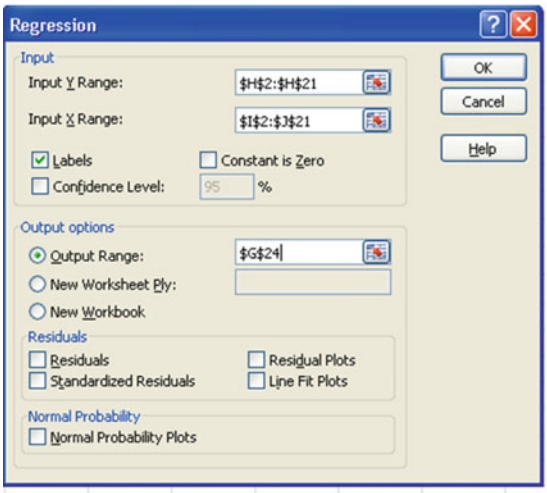
	United Technologies Corp			Wal-Mart Stores Inc			Microsoft Corp		
Year	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS
1990	47.875	5.91	1.800	42.625	1.90	0.220	76.000	2.34	0.000
1991	54.250	−8.91	1.800	33.000	1.14	0.140	68.125	2.47	0.000
1992	48.125	−0.05	1.900	53.875	1.40	0.172	70.000	2.41	0.000
1993	62.000	3.53	1.800	65.125	1.74	0.212	88.000	3.15	0.000
1994	62.875	4.40	1.900	26.500	1.02	0.130	51.625	1.88	0.000
1995	94.875	5.70	2.050	22.875	1.17	0.170	90.375	2.32	0.000
1996	66.250	3.45	1.100	20.375	1.19	0.200	120.125	3.43	0.000
1997	72.812	4.44	1.240	23.750	1.33	0.210	126.375	2.63	0.000
1998	108.750	5.36	1.390	39.812	1.56	0.270	108.375	1.83	0.000
1999	65.000	1.74	0.760	86.000	1.98	0.310	90.188	1.54	0.000
2000	78.625	3.78	0.825	54.750	1.25	0.200	80.000	1.81	0.000
2001	64.630	4.06	0.900	56.800	1.41	0.240	73.000	1.45	0.000
2002	61.940	4.67	0.980	59.980	1.49	0.280	54.700	1.45	0.000
2003	94.770	4.93	1.135	47.800	1.81	0.300	25.640	0.93	0.080
2004	103.350	5.62	1.400	53.850	2.03	0.360	28.560	0.76	0.160
2005	55.910	3.19	0.880	52.400	2.41	0.520	24.840	1.13	3.320
2006	62.520	3.81	1.015	46.110	2.68	0.600	23.300	1.21	0.340
2007	76.540	4.38	1.170	47.690	2.92	0.503	29.470	1.44	0.390
2008	53.600	5.00	1.345	50.740	3.17	0.880	27.510	1.90	0.430
2009	69.410	4.17	1.540	47.120	3.36	0.950	23.770	1.63	0.500

Source: EPS, DPS, and PPS for Dow Jones 30 are from Standard & Poor's Compustat, Wharton Research Data Services (WRDS)

Table IV.1H PPS, EPS, and DPS for Dow Jones 30industrial firms (1990–2009)

	Cisco Systems Inc			Travelers Cos Inc			Kraft Foods Inc			
Year	PPS	EPS	DPS	PPS	EPS	DPS	PPS	EPS	DPS	S&P 500
1990	25.500	1.00	0.000							541.72
1991	39.250	1.38	0.000							670.5
1992	53.375	1.33	0.000							873.43
1993	51.875	1.33	0.000							1,085.5
1994	21.000	1.19	0.000							1,327.33
1995	55.750	1.52	0.000	34.540	1.42	0.000				1,427.22
1996	51.750	1.37	0.000	35.375	1.05	0.150				1,194.18
1997	79.562	1.52	0.000	44.000	3.13	0.300				993.94
1998	95.750	1.32	0.000	31.000	3.43	0.400				965.23
1999	62.125	0.65	0.000	34.250	3.62	0.500		1.20		1,130.65
2000	65.438	0.39	0.000	42.000	4.24	1.110		1.38		1,207.23
2001	19.220	−0.14	0.000	43.970	−5.22	1.070	34.030	1.17	0.130	1,310.46
2002	13.190	0.26	0.000	14.650	0.23	0.000	38.930	1.96	0.540	1,477.19
2003	19.490	0.50	0.000	16.970	1.69	0.280	32.220	2.01	0.630	1,220.04
2004	20.920	0.73	0.000	37.070	1.56	1.055	35.610	1.56	0.745	948.05
2005	19.150	0.88	0.000	44.670	3.04	0.910	28.170	1.72	0.845	541.72
2006	17.880	0.91	0.000	53.690	6.12	1.010	35.700	1.86	0.940	670.5
2007	28.910	1.21	0.000	53.800	7.04	1.130	32.630	1.64	1.020	873.43
2008	21.990	1.35	0.000	45.200	4.90	1.190	26.850	1.24	1.100	1,085.5
2009	22.010	1.05	0.000	49.860	6.38	1.230	27.180	2.04	1.160	1,327.33

Source: EPS, DPS, and PPS for Dow Jones 30 are from Standard & Poor's Compustat, Wharton Research Data Services (WRDS)



JNJ Current Ratio

t	Y(t)	X(t-1)	Y(t-1)
1991	1.835	1.65035	1.778
1992	1.582	1.58368	1.835
1993	1.624	1.8485	1.582
1994	1.566	1.89574	1.624
1995	1.809	1.45418	1.566
1996	1.807	1.52022	1.809
1997	1.999	2.023	1.807
1998	1.364	1.93359	1.999
1999	1.771	1.99812	1.364
2000	2.164	1.83149	1.771
2001	2.296	2.0827	2.164
2002	1.683	1.94572	2.296
2003	1.71	1.95464	1.683
2004	1.962	1.84799	1.71
2005	2.485	2.09756	1.962
2006	1.199	2.44232	2.485
2007	1.51	1.96882	1.199
2008	1.649	2.03753	1.51
2009	1.82	2.18401	1.649

Fig. 16.25 Current ratio regression for JNJ

Appendix 1: Dynamic Ratio Analysis

In the Appendices, we discuss how the industry average of financial ratio can be used to do dynamic financial ratio analysis. To do the dynamic financial ratio analysis, the individual financial ratio is related to the industry average over time by a regression such as[8]

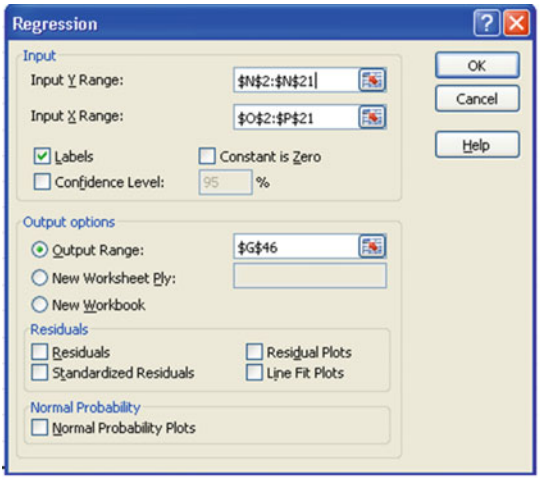
$$y_{i,t} = a_0 + a_1 x_{t-1} + a_2 y_{i,t-1} + e_t \tag{16.32}$$

where

$y_{i,t}$ = a financial ratio for ith firm in period t
x_{t-1} = industry average for a financial ration in period $t-1$
$y_{i,\,t-1}$ = a financial ratio for ith firm in period $t-1$

Use the current ratio data of both Johnson & Johnson and Merck in Table 3.9 of Chap. 3 and its industry average data during 1990–2009. Using the Microsoft Excel

[8] See Lee, Cheng F., Finnerty, Joe E.: Corporate Finance. Harcourt Brace Javanovich, San Diego (1990)



MRK Current Ratio

t	Y(t)	X(t-1)	Y(t-1)
1991	1.532	1.65035	1.332
1992	1.216	1.58368	1.532
1993	0.973	1.8485	1.216
1994	1.27	1.89574	0.973
1995	1.515	1.45418	1.27
1996	1.6	1.52022	1.515
1997	1.475	2.023	1.6
1998	1.685	1.93359	1.475
1999	1.285	1.99812	1.685
2000	1.375	1.83149	1.285
2001	1.123	2.0827	1.375
2002	1.199	1.94572	1.123
2003	1.205	1.95464	1.199
2004	1.147	1.84799	1.205
2005	1.582	2.09756	1.147
2006	1.197	2.44232	1.582
2007	1.227	1.96882	1.197
2008	1.348	2.03753	1.227
2009	1.805	2.18401	1.348

Fig. 16.26 Current ratio regression for MRK

program, we obtain regression results in terms of Eq. 16.32 as presented in Figs. 16.25 and 16.26. From the values of R^2 and t statistics associated with regression coefficients a_1 and a_2, we can conclude that the current ratio regression for both Johnson & Johnson and Merck are suitable to be used to forecast future current ratio.

Using the current ratio of 2009 for JNJ (1.82) and industry average (2.18401), we can forecast the current ratio for JNJ in 2010 as follows:

$$CR_{2010}(\text{JNJ}) = 1.9421 - 0.1189(2.18401) + 0.037(1.82) = 1.7498$$

Appendix 2: Term Structure of Interest Rate

The structure of interest rates is typically described by the yield curve. Typical yield curve diagram used to describe the relationship between yield to maturity and time to maturity term for Treasury securities. It can be shown that the following multiple regression model can be used to describe this relationship[9]:

[9] See Lee, Cheng F., Finnerty, Joseph E., Wort, Donald H.: Security Analysis and Portfolio Management. Glenview. I11, Scott, Foresman (1990), Chap. 5.

Table 16.14 Yield, time to maturity, and coupon rates for Treasury bonds and notes as of February 16, 2011

(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
R_t	t	x	R_t	t	x	R_t	t	x
0.0618	0.117808	0.875	0.5159	1.369863	0.625	1.083	2.523288	4.25
0.043	0.117808	4.75	0.5034	1.369863	4.875	1.088	2.567123	3.125
0.1088	0.2	0.875	0.5273	1.410959	1.5	1.1674	2.608219	0.75
0.1194	0.2	4.875	0.5574	1.454795	0.625	1.157	2.649315	3.125
0.1154	0.284932	0.875	0.5595	1.454795	4.625	1.2163	2.690411	0.5
0.1096	0.284932	4.875	0.5931	1.49589	1.75	1.2179	2.734247	2.75
0.1355	0.367123	1.125	0.5662	1.49589	4.375	1.2563	2.731507	0.5
0.1353	0.367123	5.125	0.6087	1.539726	0.375	1.1699	2.731507	4.25
0.1458	0.452055	1	0.5916	1.539726	4.125	1.2661	2.772603	2
0.1381	0.452055	4.875	0.6444	1.580822	1.375	1.2953	2.813699	0.75
0.1461	0.493151	5	0.6574	1.621918	0.375	1.3104	2.857534	1.5
0.167	0.536986	1	0.6369	1.621918	4.25	1.3434	2.89863	1
0.1725	0.536986	4.625	0.7001	1.663014	1.375	1.3384	2.942466	1.75
0.1866	0.619178	1	0.711	1.706849	0.375	1.3997	2.983562	1.25
0.1914	0.619178	4.5	0.6779	1.706849	3.875	1.3096	2.983562	4
0.2108	0.70411	1	0.7456	1.747945	1.375	1.3889	3.019178	1.875
0.2038	0.70411	4.625	0.7099	1.747945	4	1.4282	3.10411	1.75
0.2229	0.745205	1.75	0.7496	1.789041	0.5	1.4757	3.186301	1.875
0.2453	0.786301	0.75	0.7088	1.789041	3.375	1.5079	3.271233	2.25
0.2423	0.786301	4.5	0.7942	1.830137	1.125	1.5515	3.353425	2.625
0.2656	0.827397	1.125	0.8023	1.873973	0.625	1.5977	3.438356	2.625
0.287	0.871233	1	0.7645	1.873973	3.625	1.5634	3.479452	4.25
0.2927	0.871233	4.625	0.8251	1.915068	1.375	1.6612	3.523288	2.375
0.3065	0.912329	1.125	0.8392	1.958904	0.625	1.6947	3.605479	2.375
0.3162	0.956164	0.875	0.7789	1.958904	2.875	1.7417	3.690411	2.375
0.3211	0.956164	4.75	0.7933	2	3.875	1.6727	3.731507	4.25
0.3354	0.99726	1.375	0.8481	2.063014	2.75	1.7794	3.772603	2.125
0.3423	0.99726	4.875	0.8981	2.10411	1.375	1.8158	3.857534	2.625
0.3504	1.035616	0.875	0.8785	2.147945	2.5	1.8648	3.942466	2.25
0.3486	1.035616	4.625	0.9392	2.189041	1.75	1.8246	3.983562	4
0.3626	1.076712	1.375	0.9126	2.230137	3.125	1.778	3.983562	11.25
0.3956	1.120548	1	0.9895	2.271233	1.375	1.9036	4.060274	2.375
0.409	1.161644	1.375	0.9264	2.271233	3.625	1.9303	4.145205	2.5
0.4309	1.20274	1	0.9576	2.315068	3.5	1.9814	4.227397	2.5
0.4157	1.20274	4.5	1.0293	2.356164	1.125	1.9563	4.268493	4.125
0.4516	1.243836	1.375	1.011	2.39726	3.375	2.0309	4.312329	2.125
0.4717	1.287671	0.75	1.0691	2.438356	1	2.0703	4.394521	1.875
0.469	1.287671	4.75	1.0369	2.482192	3.375	2.112	4.479452	1.75
0.4969	1.328767	1.875	1.113	2.523288	0.75	2.0764	4.520548	4.25
1.9998	4.520548	10.625	3.2892	8.273973	3.125	4.644	27.26849	4.5
2.1709	4.564384	1.25	3.3314	8.526027	3.625	4.6873	28.00548	3.5
2.2055	4.646575	1.25	3.1969	8.526027	8.125	4.6714	28.24932	4.25
2.2467	4.731507	1.25	3.3943	8.778082	3.375	4.6665	28.24932	4.5
2.3681	5.063014	2.625	3.4429	9.030137	3.625	4.6747	28.50137	4.375

(continued)

Table 16.14 (continued)

(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
R_t	t	x	R_t	t	x	R_t	t	x
2.409	5.147945	2.375	3.3005	9.030137	8.5	4.6685	29.00822	4.625
2.4363	5.230137	2.625	3.4965	9.276712	3.5	4.6793	29.26027	4.375
2.3992	5.271233	5.125	3.3523	9.276712	8.75	4.687	29.51233	3.875
2.3637	5.271233	7.25	3.5763	9.528767	2.625	4.6832	29.76438	4.25
2.4581	5.315068	3.25	3.4047	9.528767	8.75	4.6702	30.01644	4.75
2.4865	5.39726	3.25	3.616	9.780822	2.625			
2.5263	5.482192	3.25	3.6193	10.03288	3.625			
2.488	5.523288	4.875	3.4958	10.03288	7.875			
2.5655	5.567123	3	3.5484	10.27671	8.125			
2.5921	5.649315	3	3.5904	10.52603	8.125			
2.6241	5.734247	3.125	3.6425	10.77808	8			
2.5895	5.775342	4.625	3.7672	11.52603	7.25			
2.5587	5.775342	7.5	3.7995	12.27397	7.625			
2.666	5.816438	2.75	3.856	12.52603	7.125			
2.6863	5.90137	3.25	3.9425	13.02192	6.25			
2.7073	5.986301	3.125	4.035	14.27671	7.5			
2.6763	6.027397	4.625	4.0571	14.52877	7.625			
2.734	6.063014	3	4.1296	15.02466	6.875			
2.7579	6.147945	3.25	4.2027	15.52877	6			
2.793	6.230137	3.125	4.2129	15.52055	6.75			
2.7655	6.271233	4.5	4.2404	15.76986	6.5			
2.6806	6.271233	8.75	4.2513	16.02192	6.625			
2.8329	6.315068	2.75	4.2889	16.51781	6.375			
2.8672	6.39726	2.5	4.3119	16.76986	6.125			
3.0019	6.90137	2.75	4.3724	17.52055	5.5			
3.0288	6.986301	2.625	4.3924	17.7726	5.25			
3.0062	7.027397	3.5	4.3993	18.02466	5.25			
3.0511	7.271233	3.875	4.3882	18.52055	6.125			
2.9049	7.271233	9.125	4.4112	19.26575	6.25			
3.1047	7.523288	4	4.4605	20.02192	5.375			
3.1662	7.775342	3.75	4.6037	25.02466	4.5			
2.9979	7.775342	9	4.6089	26.0274	4.75			
3.2507	8.027397	2.75	4.6023	26.27397	5			
3.0691	8.027397	8.875	4.6436	27.02466	4.375			

$$\ln(y_{it}) = a_0 + a_1 x_{1t} + a_2 x_{2t} + a_3 x_{3t} + e_t \tag{16.33}$$

Where

y_{it} = yield to maturity for ith bond to be mature in period t
x_{1t} = time to maturity for ith bond
$x_{2t} = 1/x_{it}$
x_{3t} = coupon rate for ith bond

Table 16.12 Regression results for Eq. 16.33

Regression output: $\ln(1+R_t) = a_0 + a_1(t) + a_2(1/t) + a_3(x) + e_t$			
Constant (a_0)		0.673866	
Standard error of Y estimate		0.034919	
R^2		0.76818	
Number of observations	223		
Degrees of freedom	219		
	a_1	a_2	a_3
X coefficient(s)	0.042615	−0.17833	0.03791
Standard error of coefficients	0.002674	0.017143	0.007759

Table 16.13 Estimated yield of a two-year, 3.875 % coupon note

$$\ln(1+R_t) = 0.673688 + 0.042615\,(2) - 0.17833\,(1/2) + 0.03791(3.875)$$
$$= 0.8168$$

To transform $\ln(1+R_t)$ into the yield to maturity, take the exponential of both sides of the equation:

$$R_t = \exp[0.8168]\text{-}1$$
$$= 1.2632 \text{ or } 126.32\ \%$$

Using the data of Treasury bonds and notes as reported in *The Wall Street Journal* of February 16, 2011, in Table 16.14, Lee et al. estimated Eq. 16.33 and the result is presented in Table 16.12.

Using the information in Table 16.12, the estimated yield of a two-year, 3.875 % compound rate is 126.3 % as presented in Table 16.13

Part V
Selected Topics in Statistical Analysis for Business and Economics

In the previous 16 chapters of this book, we have studied descriptive statistics, probability and important distributions, statistical inference based on samples, and regression and correlation analyses. In this last part of the book, we discuss the application of selected statistical methods in business and economics. These methods include nonparametric statistics, time-series analysis and forecasting, index numbers and stock market indexes, sampling surveys, and statistical decision theory.

Chapter 17
Nonparametric Statistics

Chapter Outline

Key Terms

Parametric test	Mann–Whitney *U* test
Classical test	Wilcoxon rank-sum test
Nonparametric test	*U* statistic
Distribution-free test	Kruskal-Wallis test
Nonparametric statistics	Spearman rank correlation test
Matched-pairs sign test	Runs test
Wilcoxon matched-pairs signed-rank test	Run
Ranks	Number of runs
Wilcoxon's *W* statistic	Mean absolute relative prediction error

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_17,

17.1 Introduction

In previous chapters, we discussed alternative tests of hypotheses. These tests were generally concerned with statistical measures such as the mean, variance, or proportion of a population. A mean, variance, or proportion is referred to as a parameter in statistics. To test these parameters, we generally assume that the sample observations were drawn from a normally distributed population. The assumption of normality is especially critical when the sample size is small. Tests such as the Z, *t,* and *F* tests discussed in Chap. 11 depend on assumptions about the parameters of the population, so all these tests are *parametric tests* or *classical tests*. A parametric test is generally a test based on a parametric model.

Recently, a number of useful hypothesis-testing techniques that do not make restrictive distribution assumptions about the parameters of the population have been developed. Such testing procedures are referred to as *nonparametric tests* or *distribution-free tests*. Distribution-free tests are valid over a wide range of distributions of populations. (However, these nonparametric tests do require certain assumptions, such as independent sample observations.) For example, a sample is taken to test the effectiveness of a new toothpaste in reducing plaque. Samples of 10 people using a new brand and 10 people using the leading brand are compared. If the distribution of plaque reduction is skewed, a test based on the assumption of normality is no longer appropriate, and a nonparametric approach is necessary. Furthermore, the nonparametric test can be used to reduce the effect of outliers.

The main advantage nonparametric tests offer is that they do not require us to assume that the sample observations were drawn from a normal distribution. In addition, nonparametric tests are easier than parametric tests to conduct and to understand. The main disadvantages of nonparametric tests are that they ignore a certain amount of information and that they are not so efficient as parametric tests.

The main purpose of this chapter is to introduce some additional *nonparametric statistics* and explore their applications in testing hypotheses. Actually, in Chap. 12, we discussed two nonparametric methods for hypothesis testing: the chi-square test for goodness of fit and the chi-square test for independence.[1] This chapter focuses on the development and use of six more nonparametric tests:

1. The matched-pairs sign test
2. The Wilcoxon matched-pairs signed-rank test
3. The Mann–Whitney *U* test (rank-sum test)
4. The Kruskal–Wallis test
5. The Spearman rank correlation test
6. The number-of-runs test

After discussing these nonparametric methods, we will examine five examples of the use of nonparametric statistical methods in business decision making.

[1] The first test is concerned with how well a set of data fits a hypothesized probability distribution. The second seeks to determine whether a relationship exists between two variables. These two tests are generally large-sample tests.

Table 17.1 Assessing political party preferences

(1) Economist	(2) Score for Democrat	(3) Score for Republican	(4) Sign of difference
1	8	6	+
2	9	4	+
3	9	6	+
4	4	5	−
5	7	8	−
6	9	3	+
7	8	5	+
8	9	6	+
9	7	7	0
10	9	8	+
.	.	.	.
.	.	.	.

17.2 The Matched-Pairs Sign Test

We begin our discussion of nonparametric tests with one of the easiest to employ, the *matched-pairs sign test*. The sign test is used to test the central tendency of a population distribution and is most frequently employed in analyzing matched-pairs data. A sign test uses the sign of the difference between two numbers rather than the actual quantitative measurements.

We will illustrate the matched-pairs sign test in terms of data obtained from a sample survey. Table 17.1 shows some of the data derived from a survey that sought to determine whether economists believed a Democratic president or a Republican president would have a more positive effect on the economy. Prior to a presidential election, 55 economists were surveyed and asked to rank, on a scale from 1 to 10, the likelihood that either a Democratic or a Republican president would have a positive impact on the economy.

Columns (2) and (3) in the table show the economists' rankings of the potential for a chief executive from each of the political parties having a positive impact on the economy; 10 represents the greatest positive impact. The last column indicates only the sign of the difference, either + or –. If there is no difference between the rankings, a 0 is displayed. A plus sign means a higher numerical score was assigned to the Democratic presidential candidate than to the Republican candidate, a minus sign means the reverse, and a zero denotes a tie score.

The null hypothesis of our test is that there is no tendency to prefer one political party over the other in assessing the president's potential impact on the economy. To implement this hypothesis test, we compare only the numbers for economists who have a preference for one political party. Hence, we do not include, in our test, data for economists who predicted that both political parties would do equally well.

Among the 55 economists surveyed, 33 stated that a president from the Democratic Party would have the greater positive impact on the economy, 17 stated that a

Republican president would have the greater impact, and 5 stated that the two parties would do equally well. Because tied cases are excluded in a sign test, our analysis will include 33 plus signs and 17 minus signs.

We want to test the null hypothesis of no difference in the impact on the economy by the political parties in question; that is, we want to test the hypothesis that plus and minus signs are equally likely to occur. We would expect an equal number of plus and minus signs if the null hypothesis were true. On the other hand, either too many pluses or too many minuses will be grounds for rejection of the null hypothesis. If we use p^* to denote the probability of obtaining a plus sign, we can state the hypotheses as:

H_0: There are no differences in the parties' impact on the economy. ($p^* = .50$)
H_1: There are differences in the parties' impact on the economy. ($p^* \neq .50$)

We use the large-sample method of the normal approximation to the binomial distribution (see Chap. 7). If the observed proportion of plus signs is $\hat{p}$, then the mean and standard deviation of the sampling distribution of $\hat{p}$ are

$$\mu_{\hat{p}} = p^*$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p^* q^*}{n}} \tag{17.1}$$

Our sample consists of 33 plus signs and 17 minus signs, so we substitute $n = 50$ into Eq. 17.1. This yields

$$p^* = .5$$

and

$$\sqrt{p^* q^*/n} = \sqrt{(.5)(.5)/50} = .071$$

Because our sample consists of 50 observations and the observed proportion of plus signs is $\hat{p} = 33/50 = .66$, our test statistic, Z, can be approximated by a standard normal distribution.

$$Z = \frac{\hat{p} - p^*}{\sqrt{p^* q^*/n}} = \frac{.66 - .50}{.071} = 2.254$$

Hence, assuming that we test the hypothesis at the 5 percent level of significance, we would reject the null hypothesis if $Z < -1.96$ or $Z > 1.96$. Accordingly, our results dictate that we reject the null hypothesis that plus and minus signs are equally likely to occur. That is, because the number of plus signs is greater than the number of minus signs, we conclude that the economists in the sample believe that a president from the Democratic Party is more likely than a Republican president to have a positive influence on the economy.

As we noted in Chap. 11, the critical interval estimate for p^* (the true proportion of positive signs), rather than the Z-value, can be used to do the null hypothesis test. The critical interval estimates for p^* are

$$\hat{p} + 1.96\sqrt{p^*q^*/n} = .66 + (1.96)(.071) = .521$$
$$\hat{p} - 1.96\sqrt{p^*q^*/n} = .66 - (1.96)(.071) = .799$$

That is, $.521 < p < .799$. This interval does not contain $p^* = .5$, so we reject the null hypothesis.

17.3 The Wilcoxon Matched-Pairs Signed-Rank Test

The *Wilcoxon matched-pairs signed-rank test* is preferable when the differences between the matched pairs can be quantitatively determined, rather than merely assigned signs. In other words, the Wilcoxon test provides a method of incorporating information about the relative size of the differences between the matched pairs in terms of *ranks*.

To illustrate how to employ the Wilcoxon matched-pairs signed-rank test, we will use a sample of net income figures for Lawrence Inc., which we assume to be emerging from a major corporate reorganization (see Table 17.2). Data are given for the 10 market regions in which Lawrence sells its product. Columns (2) and (3) of Table 17.2 show the net income figures (in millions of dollars) for each region for the year before corporate restructuring and the year after, respectively.

As in the sign test we conducted in Sect. 17.2, we calculate the difference in net income for each region before and after the reorganization. This difference is entered in column (4). Next, we determine the absolute values of the differences for each region and rank the regions accordingly from 1 to n, where n is the number of regions in our example. The smallest absolute difference is assigned the rank 1. When the difference within a particular region is 0, no ranking is assigned. Hence, in our example, the data for region 5 are no longer included in our test. When absolute differences are tied, the mean rank value is assigned to those differences. In our example, because regions 6 and 10 are tied for the rank of 2, both are assigned the rank of 2.5, which is the average of 2 and 3. These ranks are entered in column (6) or column (7), depending on their sign (+ or –) in column (4). Positive-signed ranks are listed in column (6), negative-signed ranks in column (7). Again, because the difference from region 5 is 0, only 9 samples need be included in our test.

Our table also displays the sum of the ranks in columns (6) and (7). These sums are critical to our null hypothesis, which is[2]

[2] Technically, this is not a null hypothesis, because it is stated in sample—not population—terms.

Table 17.2 Net income figures for Lawrence Inc.

(1) Market region	(2) Net income before reorganization	(3) Net income after reorganization	(4) Difference, $d = (3) - (2)$	(5) Rank of $\mid d \mid$	(6) Signed +	(7) Signed –
1	41	62	21	7	7	
2	34	49	15	6	6	
3	43	39	−4	4		4
4	29	28	−1	1		1
5	55	55	0			
6	63	66	3	2.5	2.5	
7	35	47	12	5	5	
8	42	72	30	9	9	
9	57	84	27	8	8	
10	*45*	*42*	*−3*	*2.5*	*−37.5*	*2.5* 7.5

$$\begin{aligned} H_0 &: \text{Sum of plus ranks} = \text{sum of minus ranks} \\ &\textit{Sigma}\,\text{rank}(+) = \Sigma\ \text{rank}(-) \end{aligned} \tag{17.2}$$

In other words, the null hypothesis implies that the population of positive and negative differences is distributed around the mean of zero. The test statistic, referred to as *Wilcoxon's W statistic*, is the smaller of the sum of the plus ranks (W^+) and the sum of the negative ranks (W^-):

$$\begin{aligned} W^+ &= \sum_{i=1}^{n} R_i^+ \\ W^- &= \sum_{i=1}^{n} R_i^- \end{aligned} \tag{17.3}$$

For samples of $n \leq 20$, we use Table A11 in Appendix A to obtain the critical values of the test statistic W. Note that Table A11 represents the maximum value that W can have and still be considered significant at various levels of significance.

In our example, because one observation of the difference is 0, the effective sample size $n = 10 - 1 = 9$. From Table 17.2, we obtain $W^+ = 37.5$ and $W^- = 7.5$. The two-tailed value in Table A11 corresponding to $n = 9$ and $\alpha = .05$ is 6. Consequently, we are to accept H_0 if $W^- \geq 6$. Because the value of W^- is larger than 6, we cannot reject H_0 and must conclude that the net income levels before and after corporate reorganization do not differ significantly.

Kruskal and Wallis have shown that when n is large (at least 25), W is approximately normally distributed with mean μ_w and standard deviation σ_w defined as follows[3]:

$$\mu_W = n(n+1)/4$$

[3] See Kruskal W.H., Wallis W.A.: Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. **47**(152), 583–621 (1952)

```
MTB > READ INTO C1 C2
DATA> 41 62
DATA> 34 49
DATA> 43 39
DATA> 29 28
DATA> 55 55
DATA> 63 66
DATA> 35 47
DATA> 42 72
DATA> 57 74
DATA> 45 42
DATA> END
     10 rows read.
MTB > SUBTRACT C2 FROM C1, PUT INTO C3
MTB > WTEST C3

Wilcoxon Signed Rank Test

Test of median = 0.000000 versus median not = 0.000000

                N for   Wilcoxon           Estimated
          N     Test  Statistic        P     Median
C3       10        9        7.5    0.086     -8.500
MTB > WTEST OF CENTER=1 USING C3;
SUBC> ALTERNATIVE=1.

Wilcoxon Signed Rank Test

Test of median = 1.000 versus median > 1.000

                N for   Wilcoxon           Estimated
          N     Test  Statistic        P     Median
C3       10        9        5.0    0.984     -8.500
MTB > WTEST OF CENTER=1 USING C3;
SUBC> ALTERNATIVE=-1.

Wilcoxon Signed Rank Test

Test of median = 1.000 versus median < 1.000

                N for   Wilcoxon           Estimated
          N     Test  Statistic        P     Median
C3       10        9        5.0    0.022     -8.500
MTB > WTEST OF CENTER=1 USING C3.

Wilcoxon Signed Rank Test

Test of median = 1.000 versus median not = 1.000

                N for   Wilcoxon           Estimated
          N     Test  Statistic        P     Median
C3       10        9        5.0    0.044     -8.500
```

Fig. 17.1 MINITAB output of Table 17.2

$$\sigma_W = \sqrt{n(n+1)(2n+1)/24} \tag{17.4}$$

This implies that we can compute $Z = (W-\mu_W)/\sigma_w$ and perform the standard Z test, which we examined in detail in Chap. 11. Note that the power of the signed-rank test discussed in this section is higher than that of the sign test discussed in the last section. The efficiency of the matched-pairs sign test compares to that of the Wilcoxon matched-pairs signed-rank test as $3/\pi$ to $2/\pi$. Finally, note that the t test rather than the Z test is appropriate when the sample size is smaller than 25.

The MINITAB output of Table 17.2 is shown in Fig. 17.1. To discuss the four tests presented in Fig. 17.1, we will rewrite the null hypothesis of Eq. 17.2 as

$$H_0: \text{ The population differences are centered at } d_0. \tag{17.2'}$$

The first test is to test $d_0 = 0$, and its p-value is .086. Thus, it is not significant at $\alpha = .05$, which is identical to what we found before. Both the second and third tests are one-tailed tests. Their alternative tests are $d_Q > 1$ and $d_0 < 1$, respectively. From $p = .984$ and $p = .022$, we conclude that only the third test is significant at $\alpha = .05$. The fourth test is to test $d_0 = 1$. It is significant at $\alpha = .05$. From the third and fourth tests, we conclude that the net income increases by at least one million after the company is reorganized.

17.4 Mann–Whitney *U* Test (Wilcoxon Rank-Sum Test)

We will now consider another nonparametric technique that involves comparing data from two samples. The *Mann–Whitney U test*, also referred to as the *Wilcoxon rank-sum test*, tests whether two independent samples have been drawn from two populations that have the same relative frequency distribution. Unlike a sign test, the Mann–Whitney U test directly considers the rankings of the observations in each sample.

To illustrate the procedure for the Mann–Whitney U test, we will refer to Table 17.3, which shows the research and development expenditures of 15 companies in each of two major industries, A and B.

The first step in performing the Mann–Whitney U test is to combine the two samples. In Table 17.4, we rank the firms according to the dollar value of the expenditures, 1 representing least R&D expenditure and 30 representing greatest R&D expenditure. Note that firms continue to be designated by industry in Table 17.4. The next step is to sum the ranks of the sample observations listed in Table 17.4.

Referring to the data for industry A as sample 1, we can calculate the sum of ranks of items in sample 1, designated $R_1 = 166$, as shown in the second column of Table 17.3. We designate the sum of ranks of items in sample 2 (the data for

Table 17.3 R&D expenditures of two major industries (in millions of dollars)

Rank for		Rank for	
Industry A	Industry A	Industry B	Industry B
40	1	54	8
41	2	52	6
43	3	69	15
46	4	70	16
47	5	71	17
53	7	72	18
55	9	73	19
56	10	76	20
61	11	77	21
63	12	78	22
64	13	82	25
68	14	83	26
79	23	84	27
80	24	88	29
85	*28*	89	*30*
	166		299

industry B) as R_2. Accordingly, $R_2 = 299$, as indicated in the last column of Table 17.3. In general, if several variables are tied, then we assign each the average of the ranks.

If the null hypothesis is true—in other words, if the samples from the two industries were drawn from the same population—then we would expect that the totals of these two ranks (R_1 and R_2) would be approximately equal. In order to test this hypothesis, we calculate a *U statistic*. The U statistic is a test statistic that depends on the number of observations in the samples as well as on the total of the ranks for one of the samples—in this case, R_1:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{17.5}$$

where n_1 is the number of observations in sample 1 and n_2 is the number of observations in sample 2. This test statistic could also be stated in terms of R_2:

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \tag{17.6}$$

The U statistic measures the difference between the ranked observations of the two samples and provides evidence on the difference between their population distributions. Either very small or very large U values provide evidence of the separation of the matched observations of the two samples.

Table 17.4 Ranking by dollar value of expenditure

Rank	R&D expenditure	Industry
1	40	A
2	41	A
3	43	A
4	46	A
5	47	A
6	52	A
7	53	B
8	54	A
9	55	A
10	56	A
11	61	B
12	63	A
13	64	A
14	68	A
15	69	B
16	70	B
17	71	B
18	72	B
19	73	B
20	76	B
21	77	B
22	78	B
23	79	A
24	80	A
25	82	B
26	83	B
27	84	B
28	85	A
29	88	B
30	89	B

It can be shown that the sample distribution of U has the following mean and standard deviation:

$$\mu_U = \frac{n_1 n_2}{2} \tag{17.7}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{17.8}$$

Moreover, when the numbers of observations in both samples n_1 and n_2 are in excess of approximately 10 observations, the sampling distributions of ranking dollar value of expenditure approach a normal distribution.

For our example, the U statistic, the mean, and the standard deviation are calculated as follows:

$$U = (15)(15) + \frac{15(15+1)}{2} - 166 = 179$$

$$\mu_U = \frac{(15)(15)}{2} = 112.5$$

$$\sigma_U = \sqrt{\frac{(15)(15)(15+15+1)}{12}} = 24.11$$

Using this information, we can calculate the standardized normal variate:

$$Z = \frac{U - \mu_U}{\sigma_U} = \frac{179 - 112.5}{24.11} = 2.76$$

Again, assuming a two-tailed test at the 5 percent level of significance, we find from Table A3 in Appendix A that the critical value for Z is 1.96. Because our calculated Z-value exceeds the critical value, we can reject the null hypothesis that the sample observations were drawn from the same population.

A MINITAB solution for this example is shown in Fig. 17.2. The Mann–Whitney statistic is denoted as W, which is the same as the first sample ranks. R_2 is obtained by using the identity

$$R_1 + R_2 = \frac{n(n+1)}{2}$$

where $n = n_1 + n_2$. The values of U_1 and U_2 are then easily calculated. The p-value of .0062 in Fig. 17.2 indicates that we should reject the null hypothesis that the two samples are equal at $\alpha = .05$.

When samples n_1 and n_2 are both ≤ 10, Table A12 in Appendix A may be used to obtain the critical values of test statistic R_1 for both one- and two-tailed tests at various levels of significance. The producer commodity price indexes for January 1985 and January 1986 for 6 product categories listed in the table are used to show how the Wilcoxon rank-sum test can be performed (data from *Standard & Poor's Statistical Service, Current Statistics,* Jan. 1987, pp. 12–13). Combined ranks are shown in parentheses.

Product category	January 1985	January 1986
Processed poultry	198.8 (4)	192.4 (3)
Concrete ingredients	331.0 (8)	339.0 (9)
Lumber	343.0 (10)	329.6 (7)
Gas fuels	1,073.0 (12)	1,034.3 (11)
Drugs and pharmaceuticals	247.4 (5)	265.9 (6)
Synthetic fibers	157.6 (2)	151.1 (1)

```
MTB > READ C1 C2
DATA> 40 54
DATA> 41 52
DATA> 43 69
DATA> 46 70
DATA> 47 71
DATA> 53 72
DATA> 55 73
DATA> 56 76
DATA> 61 77
DATA> 63 78
DATA> 64 82
DATA> 68 83
DATA> 79 84
DATA> 80 88
DATA> 85 89
DATA> END
     15 rows read.
MTB > MANN-WHITNEY C1 C2

Mann-Whitney Confidence Interval and Test

C1          N =  15     Median =       56.00
C2          N =  15     Median =       76.00
Point estimate for ETA1-ETA2 is      -17.00
95.4 Percent CI for ETA1-ETA2 is (-28.00,-6.00)
W = 166.0
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.0062
```

Fig. 17.2 MINITAB output of Table 17.3

From the combined ranks information in the table, we find

$$R_1 = 4 + 8 + 10 + 12 + 5 + 2 = 41$$

$$R_2 = 3 + 9 + 7 + 11 + 6 + 1 = 37$$

As a check on the ranking procedure, by substituting $R_1 = 41, R_2 = 37$, and $n = 12$ into the identity discussed previously, we obtain

$$41 + 37 = \frac{12(12+1)}{2} = 78$$

Because both n_1 and $n_2 \leq 10$, Table A12 is used to obtain the critical value R_1 statistic. With $n_1 = 6$ and $n_2 = 6$, we observe (Table A12) that at the .05 level of significance, the lower and upper critical values for the two-tailed test are, respectively, 26 and 52. Because the observed value of the test statistic $R_1 = 41$ falls between the critical values, the null hypothesis cannot be rejected. In other words, the probability distribution of economic indexes did not change during the period of January 1985 and January 1986.

Table 17.5 Calculations for the Kruskal–Wallis test scores and ranks classified by size of firm

Large		Medium		Small	
Score	*Rank*	*Score*	*Rank*	*Score*	*Rank*
79	12	69	6	83	14
96	20	78	11	66	5
86	16	85	15	51	1
88	17	62	3	94	19
76	10	63	4	71	7
91	18	73	8	61	2
81	13			74	9
$n_1 = 7$		$n_2 = 6$		$n\ 3 = 7$	
$R_1 = 106$		$R_2 = 47$		$R\ 3 = 57$	

17.5 Kruskal–Wallis Test for *m* Independent Samples

The *Kruskal–Wallis* test is a one-factor analysis of variance by ranks. It is a nonparametric test that represents a generalization of the two-sample Mann–Whitney U rank-sum test to situations where more than two populations are involved. Unlike one-factor analyses of variance (see Chap. 12), the Kruskal–Wallis test makes no assumptions about the population distribution.

This test is based on a test statistic calculated from ranks established by pooling the observations from c independent simple random samples (where $c > 2$). The null hypothesis is that the populations are identically distributed or, alternatively, that the samples were drawn from c identical populations. Let's follow the procedure through an example.

Assume that simple samples of executive vice presidents in a certain industry were drawn from firms classified into three size categories (large, medium, and small). After being assured of the confidentiality of their replies, the 20 executives were asked to rate the overall quality of their board of directors' performance in setting general corporate policy during the past 3-year period on a scale from 0 to 100, with 0 denoting the lowest rating and 100 the highest. The scores, classified by size of firm, and the rankings of the pooled sample scores are shown in Table 17.5.

The result was the following pooled ranking, with the lowest score that was actually given represented by 1 and the highest by 20.

Score	51	61	62	63	66	69	71	73	74	76	78	79
Rank	1	2	3	4	5	6	7	8	9	10	11	12
Score	81	83	85	86	88	91	94	96				
Rank	13	14	15	16	17	18	19	20				

The Kruskal–Wallis test statistic, K, compares the variations of the ranks of the sample groups:

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{c} \frac{R_i^2}{n_i} - 3(n+1) \tag{17.9}$$

where

n_i = number of observations in the ith sample
$n = n_1 + n_2 + \ldots + n_c$ = total number of observations in the c samples
R_i = sum of the ranks for the ith sample

Table 17.5 gives the sample sizes and rank sums for each sample group. Substituting into the foregoing formula, we compute the K statistic in the present example:

$$K = \frac{12}{20(20+1)} \left(\frac{106^2}{7} + \frac{47^2}{6} + \frac{57^2}{7} \right) - 3(20+1) = 6.64$$

As a check on calculations at this point, make sure the ranks sum to $n(n+1)/2 = (20)(21)/2 = 210$; here, $106 + 47 + 57 = 210$.

It can be shown that the sampling distribution of K is approximately the same as the chi-square distribution with $v = c-1$ degrees of freedom (where c is the number of sample groups). In this example, where there are 3 sample groups, the number of degrees of freedom is $v = c-1 = 3-1 = 2$. Testing the null hypothesis at the 5 percent level of significance ($\alpha = .05$) and using Table A5 of Appendix A, we find the critical value of χ^2 to be $\chi^2_{2,\,0.05} = 5.991$. Hence, our rule for the one-tailed test is as follows:

If $K > 5.991$, reject the null hypothesis.
If $K \leq 5.991$, do not reject the null hypothesis.

Because $K = 6.64$ is greater than the critical value of 5.991, we reject the null hypothesis of identically distributed populations. Therefore, we conclude that there are significant differences by size of firm in the scores assigned by these 3 samples of executive vice presidents.

MINITAB output of the Kruskal–Wallis statistics of Table 17.5 is presented in Fig. 17.3. Note that the board of directors' performance scores are stored in $C1$, whereas column $C2$ contains the sample number of each observation (1, 2, or 3). The value of the Kruskal–Wallis statistic is called H and agrees with the previous result. The p-value associated with H is .037.

```
MTB > READ C1-C2
DATA> 79 1
DATA> 96 1
DATA> 86 1
DATA> 88 1
DATA> 76 1
DATA> 91 1
DATA> 81 1
DATA> 69 2
DATA> 78 2
DATA> 85 2
DATA> 62 2
DATA> 63 2
DATA> 73 2
DATA> 83 3
DATA> 66 3
DATA> 51 3
DATA> 94 3
DATA> 71 3
DATA> 61 3
DATA> 74 3
DATA> END
     20 rows read.
MTB > KRUSKAL-WALLIS C1 C2

Kruskal-Wallis Test

Kruskal-Wallis Test on C1

C2           N    Median   Ave Rank          Z
1            7     86.00       15.1       2.58
2            6     71.00        7.8      -1.32
3            7     71.00        8.1      -1.31
Overall     20                 10.5

H = 6.64  DF = 2  P = 0.036
```

Fig. 17.3 MINITAB output of Table 17.5

17.6 Spearman Rank Correlation Test

The *Spearman rank correlation test* is a nonparametric method of correlation designed to measure the strength of association between two sets of ranked data. As we have learned, nonparametric procedures can be useful in correlation analysis where the basic data are not available in the form of numerical magnitudes but where rankings can be assigned. If two variables of interest can be ranked in separate ordered series, a rank correlation coefficient can be computed. We will consider two different cases, the first representing perfect direct correlation between two series and the second, perfect inverse correlation.

Table 17.6 Rank correlation of mathematics learning ability with physics learning ability (perfect direct correlation)

Student	Rank in mathematics ability, X	Rank in physics ability, Y	Difference in ranks, $d = X–Y$	$d^2 = (X–Y)^2$
A	1	1	0	0
B	2	2	0	0
C	3	3	0	0
D	4	4	0	0
E	5	5	0	0
F	6	6	0	0
G	7	7	0	0
H	8	8	0	0
I	9	9	0	0
J	10	10	0	0

Table 17.6 displays data on the rankings of a simple random sample of 10 students according to learning abilities in mathematics and physics. Clearly, this represents a case in which it would be almost impossible to obtain precise quantitative measures of these abilities but in which rankings may be feasible. In rank correlation analysis, the rankings may be assigned in order from high to low, with 1 representing the highest rating, 2 the next highest, and so on, or from low to high, with 1 representing the lowest rank, 2 the next lowest, and so on. The computed rank correlation coefficient is the same regardless of the rank ordering used.

The rank correlation coefficient (r_s) is computed by the formula

$$r_s = 1 - \frac{6\Sigma\, d^2}{n(n^2 - 1)} \tag{17.10}$$

where

d = difference between the ranks for the paired observations
n = number of paired observations

The calculations of the rank correlation coefficients for the two extreme cases are shown in Tables 17.6 and 17.7. In the first table, there is a perfect direct correlation in the rankings; that is, the student who ranks highest in mathematics ability is also best in physics. In the second table, there is perfect inverse correlation; that is, the student who ranks highest in mathematics is worst in physics.

In the case of perfect correlation between the ranks, $r_s = 1$; in perfect inverse correlation, $r_s = -1$. An r_s-value of zero indicates no correlation between rankings. Tied ranks are handled in the calculations by averaging. Substituting $\sum d^2 = 0$ and $\sum d^2 = 330$ into Eq. 17.10, we obtain

$$r_s = 1 - \frac{6\Sigma\, d^2}{n(n^2 - 1)} = 1 - \frac{6(0)}{10(10^2 - 1)} = 1$$

Table 17.7 Rank correlation of mathematics learning ability with physics learning ability (perfect inverse correlation)

Student	Rank in mathematics ability, X	Rank in physics ability, Y	Difference in ranks $d = X–Y$	$d^2 = (X–Y)^2$
A	I	10	−9	81
B	2	9	−7	49
C	3	8	−5	25
D	4	7	−3	9
E	5	6	−1	1
F	6	5	1	1
G	7	4	3	9
H	8	3	5	25
I	9	2	7	49
J	10	1	9	81
				330

$$r_s = 1 - \frac{6\Sigma\, d^2}{n(n^2-1)} = 1 - \frac{6(330)}{10(10^2-1)} = -1$$

The significance of rank correlation is tested in the same way as for the sample correlation coefficient r. We compute the statistic

$$t = \frac{r_s}{\sqrt{(1-r_s^2)/(n-2)}} \tag{17.11}$$

which has a t distribution with $(n-2)$ degrees of freedom. If $r_s = .90$, then

$$t = \frac{.90}{\sqrt{(1-.81)/(10-2)}} = 5.84$$

Assuming we are using a two-tailed test of the null hypothesis of zero correlation in the ranked data of the population, the critical t at a 5 percent level of significance with 8 degrees of freedom is equal to 2.306, as indicated in Fig. 17.4. We reject the hypothesis of no rank correlation and conclude that a positive linear relationship exists between rank in mathematics learning ability and in physics learning ability.

17.7 The Number-of-Runs Test

In economics and finance, we are often interested in examining the randomness of series of data. For example, if the movements of stock prices are random over time, it is impossible to forecast future stock prices accurately. Hence, it is not possible to earn abnormal profits by using data on past stock prices. This hypothesis has come

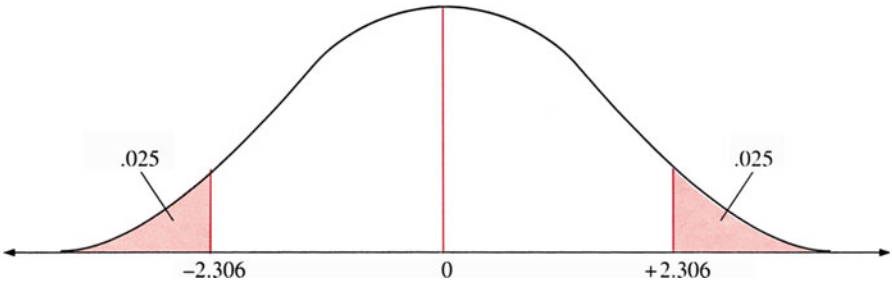


Fig. 17.4 Sampling distribution of rank correlation with $\alpha = .05$ and 8 degrees of freedom

to be known in finance as the *efficient market hypothesis*. In this section, we discuss a nonparametric procedure known as a *runs test* that can be used to examine the randomness of a series.

To test whether the collected sample data are random, we can use a nonparametric test statistic known as the *number of runs* to perform statistical inference. A *run* is a sequence of identical occurrences preceded and followed by different occurrences or by none at all. For example, suppose you toss a coin 15 times, recording each appearance of heads (H) or tails (T), and get

$$\overline{\text{HHHH}}\underline{\text{TTTT}}\overline{\text{HHHH}}\underline{\text{TTT}}$$

In this sequence, there are four runs: two runs of heads and two runs of tails.

To illustrate how a number-of-runs test works, let's consider a series of observations generated by the tossing of a coin. The probability of tossing a head or a tail is $\frac{1}{2}$, or 50 percent, so if we tossed the coin enough times, we would expect about half of the tosses to be heads and about half to be tails. However, even though this series is random, the sequence in which the head and tails appear could be HTHTHTHTHTHTHTHTHTHTHTHTHTHTHTHTHTHTH or HHHTHTTHTTTHHTTTHTHH or any of many others. The first series consists of 41 observations and 41 runs, the second of 20 observations but only 11 runs. The runs test looks at the number of runs and answers this question: Are there too many or too few runs for the series to be considered random? The hypotheses we are interested in, then, are as follows:

H_0: The series of observations are random.
H_1: The series of observations are not random.

The methods we use to do this hypothesis testing depend on whether the number of observations is smaller than 40 or, on the other hand, equal to or larger than 40. If the sample is smaller than 40, we use the Wald–Wolfowitz two-sample runs test presented in Table A10 in Appendix A.

If the number of observations is large enough (40 or more), the distribution is approximately normal, and we can use the normally distributed random variable Z defined in Eq. 17.12:

$$Z = \frac{R - \mu_R}{\sigma_R} \tag{17.12}$$

where

R = number of runs in our series
μ_R = mean value of $R = 2n_1\, n_2/n + 1$
σ_R = standard deviation of R; $\sigma_R = \sqrt{\frac{2n_1 n_2(2n_1 n_2 - n)}{n^2(n-1)}}$
n_1 = number of times we observe the first value
n_2 = number of times we observe the second value
$n = n_1 + n_2$

To illustrate how the number-of-runs test works, let's return to our previous example. Here, we generated two series of observations, by tossing a coin 41 times and 20 times, respectively:

Series 1
H T H T H T H T H T H T H T H T H T H T H
H T H T H T H T H T H T H T H T H T H T H

This series consists of 41 observations and 41 runs:

Series 2
HHH T H TT H TTT HH TTT H T HH

This series consists of 20 observations and 11 runs.

We can use Eq. 17.12 to test the randomness of the first series because $n > 40$. $R = 41$, $n_1 = 21$, $n_2 = 20$,

$$\begin{aligned}\mu_R &= 2(21)(20)/(21 + 20) + 1 \\ &= 21.49\end{aligned}$$

and

$$\begin{aligned}\sigma_R &= \sqrt{\frac{2(21)(20)[2(21)(20) - 41]}{(41)^2(41 - 1)}} \\ &= 3.16\end{aligned}$$

so Z, our test statistic, is $Z = 41 - 21.49/3.16 = 6.17$. From Table A3 in Appendix A, we find $Z_{.025} = 1.96$. Because our Z-value is 6.17, we are able to reject the null hypothesis of randomness of the series at the $\alpha = .05$ level.

Because $n < 40$ for the second series, we should use Table A10 in Appendix A to perform the test. Parts 1 and 2 of Table A10 present the critical values of the runs test at the .05 level of significance for the second series $n_1 = 10$, $n_2 = 10$, and $R = 11$. From Table A10 in Appendix A, we find that we would reject the null hypothesis at the .05 level if $R \geq 16$ or if $R \leq 6$ for the two-tailed test. Because the observed number of runs is 11, we cannot reject the null hypothesis that the series is random.

Finally, we compare the efficiency of some nonparametric tests discussed in this chapter with parametric tests as shown in the following table.

Application	Parametric test	Nonparametric test	Efficiency of nonparametric test with normal population
Two dependent samples	t test or z test	Sign test	.63
		Wilcoxon signed-ranks	.95
Two independent samples	t test or z test	Wilcoxon rank-sum	.95
Several independent samples	Analysis of variance (F test)	Kruskal–Wallis test	.95
Correlation	Linear correlation	Rank correlation	.91
Randomness	No parametric test	Runs test	No basis for comparison

From this table, we know that the efficiency of nonparametric tests is always lower than that of parametric tests if the population is distributed normally. The range of efficiency is from .63 to .95.

17.8 Business Applications

In this section, we present five applications that show how nonparametric statistics can be used in business decision making.

Application 17.1 Testing Randomness of Stock Rates of Return. The number-of-runs test can be applied to a series of stock rates of return to see whether the stock rates of return are random or exhibit a pattern that could be exploited for earning abnormal profits.

The annual data on stock rates of return for Johnson & Johnson, Merck, and the market for the period 1990–2009 are listed in Table 17.8. The numbers of runs are presented to the right of each variable. Table 17.8 indicates that the numbers of runs for rates of return for Johnson & Johnson, rates of return for Merck, and market rates of return are 11, 9, and 4, respectively.

Table 17.8 Rates of return for J&J, MRK, and S&P 500 (1990–2009)

	J&J		MRK		S&P 500	
Observations	$R_{i,t}$	*Run*	$R_{i,t}$	*Run*	$R_{m,t}$	*Run*
1	0.230		0.185		0.036	
2	0.617	1	0.879	1	0.124	
3	−0.551		−0.734		0.105	
4	−0.092	2	−0.183	2	0.086	
5	0.245		0.142		0.020	
6	0.585	3	0.754		0.177	
7	−0.410	4	0.235		0.238	
8	0.341		0.353		0.303	
9	0.288		0.410	3	0.243	
10	0.124		−0.537	4	0.223	
11	0.140	5	0.412	5	0.075	1
12	−0.431		−0.357		−0.163	
13	−0.078		−0.013		−0.168	
14	−0.021	6	−0.158		−0.029	2
15	0.249	7	−0.272	6	0.171	
16	−0.032	8	0.037		0.068	
17	0.122		0.418		0.086	
18	0.035	9	0.367	7	0.127	3
19	−0.076	10	−0.451	8	−0.174	
20	0.108	11	0.254	9	−0.223	4

If we assume that n_1 and n_2 represent "number of minus signs" and "number of plus signs," respectively, then n_1 and n_2 for all three variables are as follows:

	$R_{i,t}$ (J&J)	$R_{i,t}$ (MRK)	$R_{m,t}$ (S&P 500)
n_1	8	8	5
n_2	12	12	15

To do the test, we need to find the critical values from Table A10 in Appendix A of this book. At a 5 percent level of significance for a two-tailed test, the critical value for $n_1 = 8$ and $n_2 = 12$ is either $R \geq 16$ or $R \leq 6$; the critical value for $n_1 = 5$ and $n_2 = 15$ is $R \leq 4$. The calculated numbers of runs for $R_{\text{i,t}}$ (J&J), $R_{\text{i,t}}$ (MRK), and the market (S&P 500) rates of return are 11, 9, and 4, respectively, so we cannot reject the null hypothesis that rates of return for J&J and MRK are random. However, we can reject the hypothesis that the market (S&P 500) rate of return is random.

MINITAB output in terms of data in Table 17.8 is shown in Fig. 17.5, which indicates that all three series of data are random at $\alpha = .05$. In this MINITAB output, K represents the mean of each series. For example, the mean of the rate of return for J&J is .070. Based on the mean, we find that there are 11 observations above K and 9 observations below K. From this information, we find that there are

11 runs associated with the rate of return of J&J. Similarly, we can calculate related information for the rates of return of MRK and S&P 500.

If we assume that n_1 and n_2 represent the "number of observations above the mean" and the "number of observations below the mean," respectively, the n_1 and n_2 for all three variables are as follows:

Data Display

Row	JNJ	MRK	S&P
1	0.230108	0.185441	0.036396
2	0.616842	0.878595	0.124301
3	-0.551293	-0.733803	0.105162
4	-0.091578	-0.182990	0.085799
5	0.244915	0.142442	0.019960
6	0.584558	0.753915	0.176578
7	-0.409758	0.235280	0.237724
8	0.340804	0.352548	0.302655
9	0.287688	0.409696	0.242801
10	0.124207	-0.537078	0.222782
11	0.139719	0.411867	0.075256
12	-0.431191	-0.357447	-0.163282
13	-0.078010	-0.013313	-0.167680
14	-0.021172	-0.158294	-0.028885
15	0.248595	-0.271964	0.171379
16	-0.032496	0.036942	0.067731
17	0.122480	0.418327	0.085510
18	0.034602	0.367425	0.127230
19	-0.076435	-0.450781	-0.174081
20	0.108473	0.254065	-0.222935

Runs Test: JNJ

Runs test for JNJ

Runs above and below K = 0.0695528

The observed number of runs = 11
The expected number of runs = 10.9
11 observations above K, 9 below
* N is small, so the following approximation may be invalid.
P-value = 0.963

Fig. 17.5 MINITAB output of Table 17.8

Runs Test: MRK

Runs test for MRK

Runs above and below K = 0.0870437

The observed number of runs = 9
The expected number of runs = 10.9
11 observations above K, 9 below
* N is small, so the following approximation may be invalid.
P-value = 0.378

Runs Test: S&P

Runs test for S&P
Runs above and below K = 0.0662201
The observed number of runs = 7
The expected number of runs = 10.1
13 observations above K, 7 below
* N is small, so the following approximation may be invalid.
P-value = 0.116

Fig. 17.5 (continued)

	$R_{i,t}$ (J&J)	$R_{i,t}$ (MRK)	$R_{m,t}$ (S&P 500)
n_1	11	11	13
n_2	9	9	7

At a 5 percent level of significance for a two-tailed test, the critical values for all three variables from Table A10 are as follows:

	Lower tail ($\alpha = .025$)	Upper tail ($\alpha = .025$)
$R_{i,t}$ (J&J)	6	16
$R_{i,t}$ (MRK)	6	16
$R_{m,t}$ (S&P 500)	5	15

Figure 17.5 indicates that the observed number of runs for $R_{i,t}$ (J&J), $R_{i,t}$ (MRK), and $R_{m,t}$ (S&P 500) are 11, 9, and 7, respectively. Because all these runs are larger than lower tail and smaller than upper tail, we cannot reject the hypothesis that all three kinds of returns are random. This conclusion is consistent with that of the MINITAB output.

Application 17.2 Comparing Errors in Earnings Forecasts by Firm Size: Management Forecasts Versus Analysts' Forecasts. Here, we investigate the

Table 17.9 Comparison of forecast errors by industry

Industry	Number of forecasts	Mean absolute relative prediction error (percentage)		Wilcoxon matched-pairs signed-rank test		Level of significance (two-tailed)
		Management	*Analyst*	*Z-value*	*t-value*	
Banking and finance	12	21.8	23.0		38	.15
Utilities	22	24.8	33.3		68	.09
Manufacturing	82	30.4	29.8	−.021		.45
Chemicals	22	25.0	35.8		32	.01
Services (transportation and recreation)	18	21.8	23.0		20	.01

Source: Adapted from Jaggi, B.: Further evidence on the accuracy of management forecasts vis-à-vis analysts' forecasts. Account. Rev. **55**, 96–101 (1980)

Table 17.10 *Co*mparison of forecast errors by firm size

Firm size (revenue in millions of dollars)	Number of firms	Mean absolute relative prediction error (percentage)		Wilcoxon matched-pairs signed-rank Test significance		Level of (two-tailed)
		Management	*Analyst*	*z-value*	*t-value*	
0–99	8	43.2	37.9		9	.12
100–299	39	32.2	37.0	−2.02		.02
300–499	22	14.7	20.6		51	.02
500–999	37	27.4	28.0	1.45		.07
1000–1999	28	23.1	23.3	−1.98		.02
2000–above	22	28.2	29.8		117	.14

Source: Adapted from Jaggi, B.: Further evidence on the accuracy of management forecasts vis-à-vis analysts' forecasts. Account. Rev. **55**, 96–101 (1980)

accuracy of management forecasts relative to analysts'[1] forecasts. Jaggi (1980)[4] used sample data from 1971 to 1974 in terms of either five industries (Table 17.9) or six different firm sizes (Table 17.10). Using the Wilcoxon matched-pairs signed-rank test we discussed in Sect. 17.3, Jaggi tested the statistical differences between management forecasts and analysts' forecasts by using either *Z*-values or Wilcoxon *t*-values as indicated in both Tables 17.9 and 17.10.

The formula for calculating these two testing statistics is presented in Eq. 17.4. Jaggi used Z-values or *t*-values when the sample size is larger (or smaller) than 30. Note that *t*-values are Wilcoxon's *W* statistics defined in Eq. 17.3. In these two tables, the *mean absolute relative prediction error* (MARPE)[5] is defined as

[4] This example is adapted from results given by Jaggi, B.: Further evidence on the accuracy of management forecasts vis-à-vis analysts' forecasts. Account. Rev. **55**, 96–101 (1980)

[5] Other methods of comparing the predicted and observed values will be discussed in the next chapter.

Table 17.11 Ranking of consumer willingness to buy products made in indicated countries

Country	Order[a] for respondent (x)	Rank order for nonrespondent (y)	Difference in ranks, $d = x–y$	$d^2 = (x–y)^2$
United Kingdom	1	1	0	0
Japan	2	3	−1	1
France	3	2	1	1
Taiwan	4	4	0	0
Brazil	5	5	0	0
India	6	6	0	0
Iran	7	7	0	0
Angola	8	8	0	0
USSR	9	9	0	0
Cuba	10	10	0	0
				2

Source: Finn, D.W., et al.: An examination of the effects of sample composition bias in a mail survey. J. Mark. Res. **25**(Oct), 331–338 (1983) (Reprinted by permission of the American Marketing Association)
[a]Data collected in spring 1977

$$\text{MARPE} = \frac{|\hat{E}_t - E_t|}{E_t} \tag{17.13}$$

where $\hat{E}_t$ represents the forecast for time period t and E_t represents actual reported earnings for time period t.

Using a 5 percent level of significance, Jaggi found that only chemicals and services industries and subgroups 2, 3, and 5 are statistically significant. In other words, only in two industries and in firms of these three sizes are management forecasts statistically different from analysts' forecasts.

Application 17.3 Studying the Relationship Between Respondent and Nonrespondent in Mail Surveys. Finn et al. applied the Spearman rank correlation test to the relationship between respondent and nonrespondent in mail surveys about consumer willingness to buy products made in ten different countries (see Table 17.11).[6] Respondent rank order is obtained by averaging 387 respondent random sampling results in mail surveys. Nonrespondent rank order is obtained by a telephone interview with 10 randomly selected mail nonrespondents.

Substituting information from Table 17.11 into Eq. 17.10, we obtain the correlation coefficient:

$$r_s = 1 - \frac{6\Sigma\, d^2}{n(n^2-1)} = 1 - \frac{6(2)}{10(10^2-1)} = .988$$

[6] Finn, D.W., Wang, C.K., Lamb, C.W.: An examination of the effects of sample composition bias in a mail survey. J. Mark. Res. **25**(Oct), 331–338 (1983).

Fig. 17.6 MINITAB output of Table 17.11

```
MTB > READ INTO C1-C2
DATA> 1 1
DATA> 2 3
DATA> 3 2
DATA> 4 4
DATA> 5 5
DATA> 6 6
DATA> 7 7
DATA> 8 8
DATA> 9 9
DATA> 10 10
DATA> END
     10 rows read.
MTB > RANK C1, PUT INTO C3
MTB > RANK C2, PUT INTO C4
MTB > CORRELATION C3 C4

Correlations (Pearson)

Correlation of C3 and C4 = 0.988
```

Then, substituting $r_s = .988$ into Eq. 17.11, we obtain the t-value:

$$t = \frac{r_s}{\sqrt{(1 - r_s^2)/(n-2)}} = \frac{.988}{\sqrt{[1 - (.988)^2)]/(10-2)}} = 18.09$$

From Table A4 in Appendix A, we find $t_{.005,8} = 3.3550$. Because 18.09 is much larger than 3.3550, we conclude that rank order for respondent and rank order for nonrespondent are highly correlated. In other words, the opinions elicited from the two groups are almost identical.

MINITAB output of Table 17.11 is presented in Fig. 17.6. The correlation coefficient it shows is .988, which is identical to previous results.

Application 17.4 Testing the Randomness of the Pattern Exhibited by Quality Control Data over Time. If the process is in control, the distribution of sample values should be randomly distributed above and below the center line of a control chart, as we noted in Sect. 10.9 of Chap. 10. To test whether the pattern of, say, 10 sample observations over time appears to be random, we can use a two-tailed hypothesis test:

H_0: The sequence of sample values is random.
H_1: The sequence of sample values is not random.

Figure 17.7 shows a run of 3 consecutive points down and another run of 7 consecutive points up. Hence, there are two runs with $n_1 = 3$ and $n_2 = 7$. The total

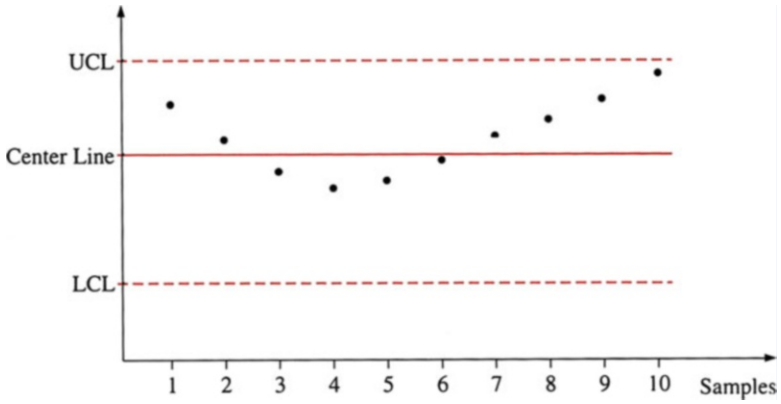


Fig. 17.7 A control chart with 10 sample observations

Table 17.12 1986 total cash compensation for 3 groups of executives

Banks and bank holding companies	Utilities	Office equipment and computers
$ 755	$520	$438
712	295	828
845	553	622
985	950	453
1.300	930	562
1,143	428	348
733	510	405
1.189	864	938

Source: Executive compensation scoreboard, *Business Week,* May 4, 1987,59–94, by special permission, copyright © 87 by McGraw-Hill, Inc

number of observations, n, is smaller than 40, so this is a small-sample case, and we can use the Wald–Wolfowitz two-sample runs-test table to do the test. Table A10 indicates that the critical value of R for $n_1 = 3$ and $n_2 = 7$ is equal to 2. The number of runs for the data presented in Fig. 17.7 is also 2. Therefore, we can reject the null hypothesis of randomness at $\alpha = 5$ percent. In other words, we conclude that the production process is not random and is out of control.

Application 17.5 Comparing Cash Compensation for 3 Different Groups of Corporate Executives. In Application 12.1 in Chap. 12, we used the analysis of variance approach to test whether there is a difference among 3 different groups of corporate executives' cash compensation. This set of data is repeated in Table 17.12.

Now, we use the Kruskal–Wallis test instead of the analysis of variance test to determine whether there is any difference in the 1986 total cash compensation for the 3 groups of corporate executives.

The MINITAB output of the Kruskal–Wallis statistic is shown in Fig. 17.8. The K statistic (H) as denned in Eq. 17.9 equals 7.80, which is significant at $\alpha = .020$.

Fig. 17.8 MINITAB output of Application 17.5

```
MTB > READ C1 C2
DATA> 755 1
DATA> 520 2
DATA> 438 3
DATA> 712 1
DATA> 295 2
DATA> 828 3
DATA> 845 1
DATA> 553 2
DATA> 622 3
DATA> 985 1
DATA> 950 2
DATA> 453 3
DATA> 1300 1
DATA> 930 2
DATA> 562 3
DATA> 1143 1
DATA> 428 2
DATA> 348 3
DATA> 733 1
DATA> 510 2
DATA> 405 3
DATA> 1189 1
DATA> 864 2
DATA> 938 3
DATA> END
     24 rows read.
MTB > KRUSKAL-WALLIS C1 C2

Kruskal-Wallis Test

Kruskal-Wallis Test on C1

C2          N     Median   Ave Rank          Z
1           8      915.0       18.1       2.76
2           8      536.5       10.5      -0.98
3           8      507.5        8.9      -1.78
Overall    24                  12.5

H = 7.80  DF = 2  P = 0.020
```

17.9 Summary

In this chapter, we discussed six nonparametric statistical tests that do not require the assumption of normality in the distribution of the population. We also gave examples of the use, in business and economic decision making, of the matched-pairs sign test, the Wilcoxon matched-pairs signed-rank test, the Mann–Whitney U test, the Kruskal–Wallis test, the Spearman rank correlation test, and the number-of-runs test.

Questions and Problems

1. Mr. John is a central New Jersey real estate salesman. He claims that the median selling price of houses in the area is about $ 100,000. To check this claim, you randomly select 10 houses that were recently sold in this area and record the following prices (in thousands of dollars).

120	115	100	113	103	97	90	111	95	88

   Using the sign test, determine whether the salesman's claim is reasonable. (Test at the .05 level of significance.)
2. Use the sign test to test, at the 5 percent level, whether the following data come from a population with median 10. Use H_1: median $\neq$ 10.

13.7	8.1	15.9	12.3	3.4	17.2	13.1	17.2
12.0	25.9	17.5	12.9	13.6	11.5	7.7	16.1
9.4	11.2	12.7	10.8				

3. Use the number-of-runs test to determine at the 5 percent level of significance whether it can be concluded that these binary sequences are not random:

   (a) 0 0 1 1 1 0 1 0 1 1 1 0 0 1 0 0 1 1 0 1 1 1
   (b) 0 0 1 1 0 0 1 1 0 0 1 1 1 1 0 0 0 0 1 0 1 0 1 0
   (c) 1 0 1 1 1 0 1 0 1 1 1 0 0 0 0 1 0 1 0 0 1 1 1 1 0 1 1 0 1 1 0 0

4. The following sequence indicates whether a manufacturer's daily production of videocassette recorders (VCRs) was above (X) or below (Y) the long-term median number of defective VCRs.
   XYYXXXYYXYYXXXYYXYXXY Y X

   (a) Does this series suggest a departure from randomness? (Test at $\alpha = .05$.)
   (b) Why would the production manager care whether this series suggested randomness?

5. The following sequence indicates whether an accident occurred at a given intersection during the rush "hour" (7:00 A.M. to 9:30 A.M). (Y indicates an accident, N no accident.)
   N Y N N Y Y N N N Y N Y Y N N N Y N Y Y N N
   (a) Is there an indication of departure from randomness at the .05 Type I error level?
   (b) Why would a transportation official be concerned about the nonrandom occurrence of accidents?

6. At the 5 percent level of significance, can we conclude that this sequence of symbols is a random series?
   + − − + + + − + + − + − + − − − + + − − + + + − − +

7. Find the Spearman rank correlation for the scores for friendliness and response time given to brand Y computers by 7 users. Test for significance, assuming that the critical value for r_s when $n = 7$ is $\pm$ .893.

User	Friendliness	Response time
1	66	79
2	75	69
3	71	84
4	61	78
5	48	65
6	90	82
7	80	90

8. In a rank correlation problem with $n = 50$ observations, the sum of the squared differences between the rank observations is $\Sigma d^2 = 12{,}500$. Calculate r.

9. The ratings assigned by a personnel manager and his assistant to several job applicants are given in the accompanying table. Use Spearman rank correlation measure r to show whether the personnel manager and his assistant disagree on how they rank the applicants. Let $\alpha = .05$.

	Applicant										
	1	2	3	4	5	6	7	8	9	10	11
Manager	9	11	10	5	3	8	2	1	4	3	2
Assistant manager	7	6	5	8	9	11	10	4	6	1	7

10. A company wishes to compare typing accuracy on two kinds of computer keyboards. Fifteen experienced typists type the same 600 words. Keyboard X_1 is used 7 times and keyboard X_2 is used 8 times, with the following results:

Number of errors								
Board X_1	13	9	16	15	10	11	12	
Board X_2	15	9	18	12	14	17	20	19

Compute the Mann–Whitney U statistic, and state the assumption for a U test. What does a large calculated U-value indicate?

11. The owner of a convenience store has noticed that the register shortages are higher at the eastside store than at the westside store. The following dollar shortages were reported for 10 randomly selected days:

Eastside store				
8.40	16.00	4.50	20.35	12.45
10.35	7.55	11.30	3.50	7.20
Westside store				
2.75	9.00	7.00	15.55	13.00
4.75	10.80	12.00	1.90	30.00

Compute the Mann–Whitney U statistic, and state the assumption for a U test. Use a 5 percent significance level.

12. The manufacturer of a new shaving cream tests 3 new advertising campaigns in a total of 21 markets. Sales in the third week after introduction are given in the accompanying table.

Shaving cream sales (*cases per thousand of population*)		
A	B	C
38	26	40
42	30	36
27	18	32
60	42	37
36	24	42
54	30	46
40	26	38

Use the Kruskal–Wallis test to determine whether the median sales levels for the 3 campaigns are different at the 5 percent level of significance.

13. Compare parametric tests to nonparametric tests. What are the assumptions of each type of test? Give some examples of each type of test.
14. The "weak form" of the efficient market hypothesis says that historical stock prices cannot be used to earn abnormal profits—that is, stock prices move randomly. Go to the library, collect 30 days of indexes from the Dow Jones Industrial Average, and use a number-of-runs test to test whether these indexes move randomly at a 10 percent level of significance.
15. Toss a coin 50 times and use a number-of-runs test to see whether the outcome of tossing a coin is random at a 5 percent level of significance.

Use the following information to answer questions 16–19. A psychologist conducting research on the differences in aptitude between males and females found 10 pairs of twins wherein one of the twins was male and the other female. Results of the tests on their math and verbal skills are shown in the table.

	Male		*Female*	
Twin	Math score	Verbal score	Math score	Verbal score
1	93	87	83	91
2	80	75	92	87
3	75	75	82	65
4	65	55	62	78
5	87	67	79	95
6	98	87	90	90
7	85	86	72	83
8	90	95	99	96
9	85	83	78	82
10	95	98	87	90

16. Use a Wilcoxon matched-pairs signed-rank test to test the hypothesis that there is a difference between the scores for males and those for females on the math portion of the test. Do a 5 percent test.
17. Use a Wilcoxon matched-pairs signed-rank test to test the hypothesis that there is a difference between the scores of males and those of females on the verbal portion of the test. Use the MINITAB program. (Hint: Follow the procedures presented in Fig. 17.1.) Do a 5 percent test.
18. Use a Wilcoxon matched-pairs signed-rank test to test the hypothesis that there is a difference between the math and verbal scores for the male twins. Do a 5 percent test.
19. Use a Wilcoxon matched-pairs signed-rank test to test the hypothesis that there is a difference between the math and verbal scores for the female twins. Use the MINITAB program. (Hint: Refer to Fig. 17.1.) Do a 5 percent test.
20. A statistics professor is interested in whether there is any difference between the scores in the first-period statistics class and those in her third-period statistics class. She collects the information shown in the accompanying table. Use a rank-sum test to test at a 10 percent level whether the scores in the first-period class and those in the third-period class are different. Use the MINITAB program. (Hint: Refer to Fig. 17.2.)

Period 1	Period 3	Period 1	Period 3
85	84	88	95
72	93	90	88
93	87	95	64
65	80	86	63
88	55	92	68
90	95	98	70
55	75	62	71
82	72	70	67
75	76	71	85
89	88	65	90
62	80	73	72
60	75	55	77

(continued)

(continued)

Period 1	Period 3	Period 1	Period 3
80	62	46	86
54	60	69	88
63	90	70	90

21. The prices for ABC Company's stock over a 12-day period follow. Test, at the $\alpha = .05$ level, whether changes in the stock price are random.

83.20	79.21	89.82
81.15	78.30	90.10
79.32	79.65	89.75
80.10	80.27	92.25

22. The following array shows price changes for silver. A + denotes an increase in price from the previous day; a – denotes a decrease in price from the previous day.
+ + + + + − − − + − + + + − − − − + − − + − + − + − −
Test, at the 10 percent level of significance, whether changes in the price of silver are random.
23. Suppose we want to test whether the number of times black or red comes up on a roulette wheel is random. In 100 spins of the wheel, we find that black comes up 48 times and red comes up 52 times. The number of runs in this sample is 48. Test, at the 1 percent level of significance, whether the appearance of red or black is random.
24. Suppose you toss a coin 50 times and receive 22 heads and 28 tails with 25 runs. Is the tossing of a head or tail random? Test this hypothesis at the 5 percent level of significance.
25. You are given the following 4 samples of price changes for a stock. Count the number of runs in each sample.

(a) + + + − − +	(c) + + + + + −
(b) + − + + − −	(d) − − + + − −

26. Suppose you collect the following information on salaries for two groups of college professors: professors in the sciences and professors in the humanities.

Science professor	Humanities professors
$52,500	$29,200
$68,270	$42,700
$55,000	$51,000
$48,900	$37,000
$75,000	$41,000

(a) Compute the ranks for each group.
(b) Do we lose information by converting numerical information into ranks?
(c) Are means and variances of ranks meaningful?

27. In a TV newsroom, 30 economists were asked one by one whether each was optimistic about the future of the economy. The answers (in order) were
+ + + + − − − − − − + + − − + − + + + − − − + + + + − − − +
The alert TV audience suspects that when the economists answered the question, each was influenced by the answer of the economist who responded right before him or her. Do a test to see whether this suspicion can be verified. Use a 5 percent test.
28. A personnel manager is considering keeping 2 out of 10 summer interns on, in a regular job, after the summer is over. Before this personnel manager accepts the subjective evaluation of the candidates, she wants to see some consensus between two evaluators. The rankings of the candidates by two different senior managers who work with them are summarized here. Do you think there is some kind of consensus between the two managers? Do a 5 percent test.

Candidate	A	B	C	D	E	F	G	H	I	J
Ranking by first manager	1	3	5	7	9	2	4	6	8	10
Ranking by second manager	2	4	6	8	10	1	3	5	7	9

29. Two senior managers were asked to score the performance of 10 job candidates in question 28. The scores assigned are believed to reflect the subjective judgments of the two senior managers, and they are believed not to follow a normal distribution. The scores obtained by the 10 candidates were as follows:

Candidate	A	B	C	D	E	F	G	H	I	J
Score by first manager	72	76	77	78	79	89	90	87	88	82
Score by second manager	71	74	83	84	85	73	92	86	73	95

Would you say the second manager is tougher in scoring? Do a 5 percent test.
30. A consumer organization wants to know whether you get what you pay for when you buy a stereo system. Its crew ranked 10 stereo systems and listed their ranks and prices in the accompanying table. Do the data support the alternative hypothesis that you get what you pay for? Do a 5 percent test.

Rank	Price
1	200
2	220
3	230
4	190
5	170
6	250
7	280
8	240
9	300
10	350

31. Two kinds of emission controls were installed and tested in 30 cars of the same make. The following table shows the test results, in the unit of emission. Can you argue that the two kinds of emission controls have a different effect? Do a 10 percent test. Assume that the data do not come from a normal distribution.

Emission control A				Emission control B	
16	17	15	18	17	16
17	15	12	19	20	11
22	21	11	18	22	9
23	22	11	19	25	27
10	10	26	11	14	12

32. The following data are the win–loss record of a professional baseball team during the last 34 games. Do you think the team is "streaky"? Do a 5 percent test. The + sign represents a win, the – a loss.
+ + – – + + – – + – + + + + – – – – + + + + – + – + + – + + + – + +

33. A market analyst stopped people in the local shop ping mall and asked them to rate two kinds of shampoo on a scale of 1–4, 1 being the lowest. The ratings of the two different brands of shampoo are listed here. Do a test to determine whether brand B is better than brand A. Use 5 percent.

Brand		*Brand*		*Brand*		*Brand*	
A	B	A	B	A	B	A	B
1	4	2	4	2	3	3	4
1	2	2	4	2	1	2	4
2	2	1	2	2	2	3	4
2	1	2	3	4	4	3	4
4	2	2	3	3	4	2	3
3	2	2	4	1	3	1	2
3	2	2	4	4	2	3	4
2	2	1	4				

Use the following information to answer questions 34–37. It is believed that American League pitchers should have higher earned-run averages because of the designated hitter rule. A baseball analyst collected 10 pitchers' earned-run averages from each of the 4 divisions in the major leagues and ranked them in the following table. (A smaller rank means a smaller earned-run average.)

National league		American league	
East	West	East	West
1	3	4	5
2	40	39	38
6	7	9	18
16	13	21	22
24	20	27	28
33	34	37	35

(continued)

(continued)

National league		American league	
East	West	East	West
8	10	11	12
15	14	17	19
23	25	26	29
30	31	32	36

34. Do a test to determine whether the American League pitchers have higher earned-run averages. Use 10 percent.
35. Do a test to determine whether the 4 different divisions have the same earned-run average. Use 10 percent.
36. Do a test to determine whether the two divisions in the National League are equal in earned-run average. Use 10 percent.
37. Do a test to determine whether the American League West has a higher earned-run average than the National League West. Use 10 percent.
38. "Pitching wins the pennant" is one of the important theories in baseball. A statistician wants to know whether this is true. He used last year's results, which are given in the accompanying table, as the sample. From the performance of these 12 teams, can he argue that winning percentage and earned-run average are negatively correlated? Do a test at a 10 percent level of acceptance, assuming that the data do not follow a normal distribution.

Winning percentage	Earned-run average
.634	3.24
.611	3.35
.598	3.36
.573	3.21
.531	3.87
.521	3.69
.479	3.98
.469	4.23
.427	4.59
.402	4.21
.389	3.72
.366	3.98

Use MINITAB and the following information to answer questions 39–42. A new production manager who believes that music can improve productivity arranges for music to be played on two out of three assembly lines. The first assembly line has no music. The second assembly line hears classical music. The third assembly line hears "easy-listening" music. The productivity in the 10 working days after the experiment began is summarized in the following table. The data are believed not to follow a normal distribution.

Assembly lines		
1	2	3
110	114	130
157	159	139
121	120	96
103	160	140
149	142	116
123	130	142
142	112	99
134	119	133
124	127	111
118	116	105

39. Can the production manager say that playing music (of whatever type) is better than not playing music? Do a 5 percent test. (Hint: Refer to Fig. 17.2.)
40. Is playing classical music better than playing easy -listening music? Do a 5 percent test. (Hint: Refer to Fig. 17.2.)
41. Some people say it makes no difference whether music is played. Can you refute this argument? Do a 5 percent test. (Hint: Refer to Fig. 17.2.)
42. Is playing classical music better than not playing any music? Do a 5 percent test. (Hint: Refer to Fig. 17.2.)
43. Assume that the rank (quality) and price of 5 stereos are as shown in the following table. Show that the Spearman rank correlation coefficients are 1 and – 1, respectively.

Price	Quality	Price	Quality
1	1	1	5
2	2	2	4
3	3	3	3
4	4	4	2
5	5	5	1

44. Is seniority related to hourly wage? A personnel manager collects data on the 10 employees' hourly wages and their seniority and ranks them in the following table. The rankings are created by assigning large ranks to higher numbers.

Hourly wage	Seniority	Hourly wage	Seniority
1	2	6	5
2	3	7	6
3	1	8	9
4	4	9	10
5	7	10	8

Do the data support the alternative hypothesis that hourly wage and seniority are related? Do a 10 percent test.

45. Can money buy a championship? A baseball writer is interested in knowing whether high salaries can produce good performance. He collected the average salaries of 12 teams and their winning percentages. The data are ranked in the accompanying table.

Ranking of winning percentage	Ranking of average salaries
1	3
2	2
3	4
4	1
5	5
6	6
7	9
8	10
9	11
10	12
11	8
12	7

Do the data support the argument that winning takes money? Do a 5 percent test.

46. The debt/equity ratio is computed by dividing total debt by total assets. It is used to measure how much leverage a firm uses. A financial analyst feels that the debt/equity ratio in industry A is higher than that in industry B. He randomly selected 20 firms from industries A and B, obtaining the following numbers.

A	B	A	B
.76	.23	.78	.67
.92	.78	.34	.23
.54	.76	.73	.24
.74	.32	.54	.34
.75	.13	.43	.22

Do a 5 percent test to decide whether industry A's debt/equity ratio is higher. Assume the data do not follow a normal distribution.

47. The New Land Food Corporation is considering retiring 10 machines. For replacement, it can order some machines of the same model or it can switch to the new model. The company has decided to try out one of the new machines before it makes a large investment in many of them. The daily productivity figures for the new machine and an old machine for the last 12 working days are recorded in the following table. Do a test to determine whether the new machine is better. Use 5 percent.

Day	1	2	3	4	5	6	7	8	9	10	11	12
New	23	21	34	24	34	33	22	32	21	15	34	33
Old	21	17	32	23	24	29	34	33	23	23	34	23

48. The dean of a business school believes that a good knowledge of economics helps business students do well in other business courses. He randomly pulled out 15 students' grades in economics and related business courses and ranked them in the accompanying table.

Economics	Business	Economics	Business
7	1	8	10
3	6	13	11
2	3	9	8
4	5	1	2
15	14	11	15
5	4	12	12
10	9	6	7
14	13		

Do a test to determine whether economics grade and grades in other business courses are related. Use 5 percent.

Use the following information to answer questions 49–53. The personnel manager wants to know whether employees strictly follow the rule for their lunch break. He suspects that those who are more senior in the company tend to take a longer lunch. He classifies the employees into three categories according to their seniority in the company. The amounts of lunch time (in minutes) are summarized in the following table. Assume these data do not follow a normal distribution.

Time with the company		
Group 1, less than 1 month	Group 2, between 1 and 6 months	Group 3, more than 6 months
43	50	63
39	55	66
42	54	59
49	49	55
50	51	57
39	60	62
47	57	53
48	55	59
39	42	56
47	38	78

49. Do the data support the hypothesis that the three different groups of employees do not spend the same amount of time at lunch? Do a 10 percent test.
50. Do the data support the hypothesis that group 1 spends less time at lunch than group 2. Do a 5 percent test.
51. Do the data support the hypothesis that group 2 spends less time at lunch than group 3. Do a 5 percent test.
52. Now assume that the data for the three different groups are actually data for the same group of individuals being observed at different seniority levels in the

company. That is, group 1 represents a sample of employees who have worked for less than 1 month; group 2 represents these same employees when they have worked for the company between 1 and 6 months. Under these circumstances, will you change the way you did questions 50 and 51?

53. Using the same assumption about the data, do a test to determine whether a person who uses more time for lunch during his first month of work will spend more time at lunch during the next 5 months at work. Use 5 percent.
54. Two experts in the insurance field rank 10 insurance companies in terms of their financial soundness. The rankings of the two experts are recorded here.

Company	A	B	C	D	E	F	G	H	I	J
Analyst I	2	3	5	7	10	1	6	4	8	9
Analyst II	3	4	7	6	8	2	10	1	5	9

Do the data support the hypothesis that there is some kind of consensus between A and B? Do a 5 percent test.

Use the following information to answer questions 55–58. A movie theater opens 3 ticket windows operated by 3 ticket sellers. If the sellers are equally effective, they should sell about the same numbers of tickets. The table gives the ticket sales of the ticket sellers for the last 10 shows. Assume the data do not follow a normal distribution.

	Seller		
Show	A	B	C
1	340	330	350
2	310	320	301
3	300	279	295
4	234	245	235
5	257	256	273
6	297	296	313
7	316	317	354
8	277	232	243
9	241	250	253
10	281	271	248

55. Consider sellers A and B. Compute the rank correlation between their sales.
56. Do a test to determine whether sellers A, B, and C are equally effective in selling tickets. Use 5 percent.
57. Do a test to determine whether sellers A and B are equally effective in selling tickets. Use 5 percent.
58. Do a test to determine whether sellers B and C are equally effective in selling tickets. Use 10 percent.
59. The plant manager in a manufacturing company wants to know whether the morning productivity is higher than the afternoon productivity. He collected the productivity numbers for 10 workers on a certain day and summarized them in

the following table. Assume that the productivity numbers do not follow a normal distribution. Do a test to determine whether morning productivity is higher. Use 5 percent.

Morning productivity	Afternoon productivity
34	35
32	33
29	31
30	36
42	41
45	33
44	42
43	39
32	31
45	40

(a) Do the test assuming that each of the 10 pairs of numbers belongs to a certain worker.
(b) Do the test assuming that we do not know to whom the 10 numbers in the morning and the 10 numbers in the afternoon belong.

60. A basketball player shot 32 times in a game, and her coach recorded the results of the 32 shots:
+ − + + − + + + + − − − − + − − − − + + + + + + − − − + − + + −
where + means "score" and − indicates "miss." The coach says that the player is "streaky." Do the data support what the coach says? Do a 5 percent test.
61. A college professor believes that those students who do well on the midterm exam tend to do well on the final. He ranked the midterm and final exams of 30 students in his class.

Midterm	Final	Midterm	Final
1	1	16	18
2	4	17	17
3	5	18	16
4	3	19	19
5	2	20	20
6	6	21	22
7	14	22	21
8	7	23	23
9	13	24	27
10	8	25	24
11	12	26	25
12	9	27	26
13	11	28	28
14	10	29	29
15	15	30	30

Do a test to see whether you can verify the professor's theory. Use 5 percent.

62. Movie critics are generally thought to be very subjective. At the end of the year, two critics rank 20 films as shown in the table.

Critic A	Critic B	Critic A	Critic B
1	5	11	11
2	3	12	10
3	1	13	12
4	6	14	14
5	2	15	18
6	7	16	19
7	8	17	15
8	4	18	20
9	9	19	17
10	13	20	16

Can you say that the two movie critics have different views? Do a 5 percent test.

63. Use MINITAB to do question 70 in Chap. 12, assuming that the data do not follow a normal distribution. Use 5 percent.

Use the following information to answer questions 64–66. As a result of the rising costs in health insurance, an insurance company decides to help its clients improve their health and cut down their bills. Thirty large companies that bought the group health insurance were picked to institute a "quit smoking" program and/or an exercise program. The amounts by which insurance claims were reduced are ranked in the following table.

"Quit smoking" program	Exercise program	Both programs
1	3	5
4	2	6
12	10	7
16	14	8
17	15	9
18	19	11
24	21	13
27	25	20
28	26	22
29	30	23

64. Can you argue that the three different groups are equally effective in reducing insurance claims? Do a 5 percent test.
65. Do a 5 percent test to determine whether the group that implemented both programs is more effective than the group that used only the "stop smoking" program.
66. Do a 5 percent test to determine whether the group using both programs is more effective than the group using only the exercise program.

67. An advertising agency tries to show the effect of a successful advertising campaign. It talked to a food company that is producing a new product, spaghetti sauce. The arrangement is to market the product in three different ways: no frills, a low-budget campaign, and a big-budget campaign. The three spaghetti sauces have the same ingredients and are sold side by side in different packages. The sales in 10 supermarkets during the first month are recorded in the following table.

No frills	Low budget	High budget
301	402	423
326	355	364
337	348	359
362	351	340
383	372	389
321	356	354
362	375	378
357	358	356
334	335	338
310	312	332

(a) Do a 5 percent test to determine whether the three approaches result in different sales. Assume the sales do not follow a normal distribution.
(b) Do a 5 percent test to determine whether the high-budget campaign is better than the low-budget campaign.

68. A statistician wants to compare the prices of beef in New York and Los Angeles. He collected beef prices from 10 supermarkets in each city and recorded the prices.

Beef price (*in dollars per pound*)			
New York	Los Angeles	New York	Los Angeles
3.05	2.95	4.21	4.09
4.35	3.33	3.72	3.85
3.37	3.45	3.29	3.65
3.42	3.50	3.95	3.89
4.05	4.32	4.95	3.76

Can you argue that beef prices are different in the two cities? Do a 5 percent test.

69. The Better Business Bureau suspects that High-Cost Gas Station is consistently charging higher prices for repairs than other garages. The BBB sent 10 cars to High-Cost for an estimate of repair costs. Then, it sent the same 10 cars to Expressway Gas Station. The results are reported here.

Car	High-Cost	Expressway
1	$678	$579
2	784	321

(continued)

(continued)

Car	High-Cost	Expressway
3	653	654
4	673	642
5	732	738
6	758	721
7	632	621
8	654	311
9	521	411
10	432	421

Can you argue that High-Cost Gas Station is charging higher prices than Expressway? Do a 10 percent test.

70. A statistics instructor is curious about whether a relationship exists between the scores of students who purchased the student workbook and those of students who did not. Because he did not require students to buy the workbook, some bought it and others didn't. The grades are recorded here and are believed not to follow a normal distribution.

Bought workbook	Did not buy workbook
75	73
74	65
80	64
81	63
82	62
77	78
76	79
72	71
70	69
67	68
83	87
84	
85	
86	

Do a test to see whether buying the workbook improved the grades. Use 5 percent. Assume that the data do not follow a normal distribution.

Use the following information to answer questions 71 and 72. A national survey was conducted to study the increase in the cost of personal health insurance. Forty people from 3 regions were asked about the increase in the table year. The results are compiled in the following table; the data do not follow a normal distribution.

Northeast	West Coast	Southeast
$320	$232	$254
279	257	231
283	264	242

(continued)

(continued)

Northeast	West Coast	Southeast
281	232	220
273	243	253
274	221	223
267	215	210
279	275	263
292	262	275
284	211	227
273	267	262
275	268	231
258	291	
252	250	

71. Do a test at the 5 percent level of significance to determine whether the average increases in the 3 regions are the same.
72. Do a test at the 5 percent level of significance to determine whether the average increase in the Northeast is higher than that on the West Coast.
73. The Dow Jones Industrial Averages from 1961 to 1986 are recorded in the following table (data from *Dow Jones Investor's Handbook* (1986), *Wall Street Journal,* January 2, 1987). Do a 5 percent test to determine whether the Dow Jones Industrial Average is a random series of data.

Year	DJIA	Year	DJIA
1961	731	1974	759
1962	652	1975	802
1963	763	1976	975
1964	874	1977	835
1965	969	1978	805
1966	786	1979	839
1967	905	1980	964
1968	944	1981	899
1969	800	1982	1,047
1970	839	1983	1,259
1971	885	1984	1,212
1972	951	1985	1,547
1973	924	1986	1,896

Use the following information to answer questions 74–76. Ron Moy has taught corporate finance for years, but this is the first year he has included the use of the computer to his course. Actually, one class did not use the computer. A second class had to use the computer. The third class was introduced to the computer but had the option not to use it. The test results that follow are grades on standardized tests and are believed not to follow a normal distribution.

Class 1	Class 2	Class 3
76	78	79
72	75	73
69	63	68
71	81	88
33	87	62
67	65	86
62	77	66
60	80	70
59	65	78
58	69	65

74. Do a test to determine whether the classes have different averages. Use 5 percent.
75. Perhaps not surprisingly, a colleague from the computer science department strongly argues for using the computer. Can you prove for him that class 2 is better than class 1? Do a 5 percent test.
76. A psychology professor argues that we should not impose any restrictions on the students. Not all students benefit from using the computer. He suggests that the best method is giving the students freedom of choice. Do a 5 percent test comparing class 3 with class 2. Do the data support the psychology professor's viewpoint?
77. A supervisor in the local factory feels that productivity during overtime is lower than productivity during normal working time. In order to verify his belief, he collected data on productivity per hour during both normal time and overtime. The productivity rates, in units per hour, are recorded in the table.

Overtime			Normal time		
249	253	257	250	251	255
252	254	256	260	259	258
261	262	265	264	266	267
263	268	269	273	274	275
270	271	272	278	277	276

Do a 5 percent test to determine whether normal-time productivity is higher than overtime productivity.

78. A psychologist believes that a person who watches the evening news one day is more likely than others to watch it the next day. To show that this is true, he kept track of his wife's TV viewing pattern without letting her know about the experiment. After 35 days of observation, he recorded the following TV viewing pattern:

WWOOOWWWWWOOOWWWWO WWWWOOOOOWWWWWWW

where W means "watch" and O means "miss." Do the data support the psychologist's theory? Do a 5 percent test.

79. A company produces 3 kinds of fruit baskets that are sold at the same price. The fruit baskets are displayed in 9 supermarkets in the local area. The sales are recorded in the following table.

	Baskets		
Supermarket	A	B	C
1	40	43	45
2	44	41	47
3	42	48	61
4	74	60	63
5	73	59	64
6	72	75	58
7	69	70	57
8	68	54	67
9	49	52	53

Do a 5 percent test to determine whether the 3 different baskets are equally popular. Assume the data do not follow a normal distribution.

80. An old proverb in baseball is "pitching wins pennants." Seven teams' pitching, in terms of earned-run average (ERA), and their standing in the American League West are given in the following table (data from *Associated Press,* June 17, 1990).

Team	Standing	ERA
Oakland	1	2.94
Chicago	2	3.01
Minnesota	3	4.16
California	4	3.49
Seattle	5	3.90
Texas	6	4.32
Kansas City	7	4.01

Do a 5 percent test to determine whether there is a relationship between ERA and team standing.

81. Use the Spearman rank correlation test to investigate the relationship between market rates of return, in terms of the S&P 500, and the annual rate of return on 3-month Treasury bills during 1970–1990, which are indicated in the following table.

	Rate of return	
Year	3-month T-bill	S&P 500
1970	6.46	.0010
1971	4.35	.1080
1972	4.07	.1557
1973	7.04	−.1737
1974	7.89	−.2964

(continued)

(continued)

	Rate of return	
Year	3-month T-bill	S&P 500
1975	5.84	.3149
1976	4.99	.1918
1977	5.27	−.1153
1978	7.22	.0105
1979	10.04	.1228
1980	11.51	.2586
1981	14.03	−.0994
1982	10.69	.1549
1983	8.63	.1706
1984	9.58	.0115
1985	7.48	.2633
1986	5.98	.1462
1987	5.82	.0203
1988	6.69	.1240
1989	8.12	.2725
1990	7.51	−.0656

82. Money magazine (Money, March 2003) reports percentage returns and expense ratios for top bond funds under 4 categories: US government (G), high-yield corporate (H), tax-exempt (T), and world bond funds (W). Can we conclude that there is significant difference in the expense ratio among the 4 types of bond funds? Do a Kruskal–Wallis test at 5 percent.

G	H	T	W
5.0	9.7	5.6	4.5
4.9	8.8	5.1	4.2
4.5	7.6	4.5	7.4
3.6	7.1	3.0	8.8
3.9	7.1	4.5	3.4
4.4	8.0	3.6	4.0
4.5	9.7	5.0	4.4
4.9	8.4	4.2	3.7

83. A random sample consists 10 young and 10 old investors. Their earnings in investment in thousand dollars are as follows:

Young	30	37	41	41	42
Old	45	48	50	52	52
Young	43	45	48	48	49
Old	54	55	58	61	64

Use Mann–Whitney U test to decide whether the young investors earn significantly different from the old investors for $\alpha = 0.05$.

84. Two financial analysts rated 10 mutual funds from 1 to 100 according to funds' recent performances. The higher the rating is, the better the efficacy. The results of their ratings are as follows:

	A	B	C	D	E	F	G	H	I	J
Analyst 1	60	40	90	10	20	70	30	80	50	85
Analyst 2	50	60	95	20	30	80	10	70	40	90

Calculate the Spearman rank correlation.

85. (Problem *84* continued.) Is there a positive rank correlation between the rankings of them? $\alpha = 0.05$.

86. (Problem *84* continued.) Use a Wilcoxon signed-rank test to test the hypothesis that there is a difference between the ratings of analyst 1 and those of analyst 2.

Chapter 18
Time Series: Analysis, Model, and Forecasting

Chapter Outline

Key Terms

Time-series data	Lagging indicators
Cross-section data	Exponential smoothing
Trend component	Exponential smoothing constant
Seasonal component	Exponential smoothing
Cyclical component	Mean squared error
Irregular component	Holt–Winters forecasting model
Percentage of moving average	Autoregressive forecasting model
Seasonal index	Trend–cycle component
Seasonal index method	Trading-day component
Leading indicators	
Coincident indicators	

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_18, © Springer Science+Business Media New York 2013

18.1 Introduction

In the first 17 chapters of this book, we used both time-series and cross-sectional data to show how statistical analysis techniques can be used in economic and business decision making.

Time-series data are any set of data from a quantifiable (or qualitative) event that are recorded *over time*. For example, we read newspapers every day and can obtain the Dow Jones Industrial Average (DJIA) index over time. The series of DJIA index values, ordered through time, constitutes time-series data. Other types of time-series data are based on the rate of inflation, the consumer price index, the balance of trade, and the annual profit of a firm.

Cross-sectional data are observations made on individuals, groups of individuals, objects, or geographic areas *at a particular time*. For example, price per share for N firms in 1991 is a set of cross-sectional data. On the other hand, price per share for General Motors over time, P_t ($t = 1, 2, \ldots, T$), is a set of time-series data.

The purpose of this chapter is to describe components of time-series analyses and to discuss alternative methods of economic and business forecasting in terms of time-series data. First, a classical description of three time-series components is offered. Then the moving average and seasonally adjusted time series are explored. Time trend regression, exponential smoothing and forecasting, and the Holt–Winters forecasting model for nonseasonal series are investigated in detail. Finally, the autoregressive forecasting model is discussed in some detail. Appendix 1 addresses the Holt–Winters forecasting model for seasonal series.

18.2 The Classical Time-Series Component Model

Several factors result in the interdependence of time-series data over time; these factors are trend, seasonal, and business cycle factors. For example, the current earnings of a growing company tend to be greater than its earnings in the period just ended, and, of course, the expected earnings in the next period will be greater than the current earnings. Therefore, the correlation between any adjacent earnings is positive, and this is due to the trend factor. Seasonal factors also contribute to the interdependence of time-series data. Retail sales in the fourth quarter account for a major portion of total annual sales of department stores. This seasonal factor ensures that the sales volume in the fourth quarter of each year is highly correlated with the fourth-quarter sales volume of any other year. The business cycle is another cause of inderdependency in a time-series model. In short, it is traditionally assumed that the total variation in a time series is composed of four basic components: a *trend component*, a *seasonal component*, a *cyclical component*, and an *irregular component*. We will now discuss these four components in some detail.

Table 18.1 Earnings per share of Johnson & Johnson

Year	EPS
2001	1.87
2002	2.2
2003	2.42
2004	2.87
2005	3.5
2006	3.76
2007	3.67
2008	4.62
2009	4.45
2010	4.85

18.2.1 The Trend Component

A trend is a pattern that exhibits a tendency either to grow or to decrease fairly steadily over time. For example, the earnings per share (EPS) of Johnson & Johnson exhibit two separate trends (or a quadratic trend) over time (see Table 18.1 and Fig. 18.1). One of the trends is from 2001 to 2007, the other from 2008 to 2010.

18.2.2 The Seasonal Component

The phenomenon of seasonality is common in the business world. Retailers can rely on greater sales volume in December than in any other month; stock returns are typically higher in January than in most other months—the "January effect."

Table 18.2 and Fig. 18.2 show earnings per share of IBM Corporation over a period of 44 quarters (first quarter 2000 to fourth quarter 2010). The table offers evidence of seasonal behavior for all quarters. The fourth-quarter figures tend to be relatively high, whereas those in the first quarter are relatively low. This seasonal behavior is quite clear in Fig. 18.2, where an obvious pattern almost repeats itself each year.

18.2.3 The Cyclical Component and Business Cycles

Cyclical patterns are long-term oscillatory patterns that are unrelated to seasonal behavior. They are not necessarily regular but instead follow rather smooth patterns of upswings and downswings, each swing lasting more than 2 or 3 years. Figure 18.3 demonstrates the cyclical pattern of the S&P 500 Composite Index during the period of 2000–2010, which will be discussed in detail in the next chapter.

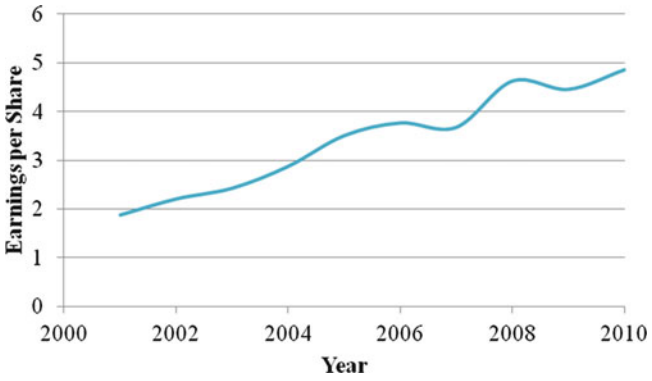


Fig. 18.1 Earnings per share of Johnson & Johnson

Table 18.2 Quarterly earnings per share of IBM Corporation

	Quarter			
Year	*1*	*2*	*3*	*4*
2000	0.85	1.95	3.06	4.58
2001	1.02	2.22	3.21	4.69
2002	0.75	1.01	2.01	3.13
2003	0.8	1.8	2.84	4.42
2004	0.81	1.84	2.77	4.48
2005	0.86	2.02	2.97	4.99
2006	1.09	2.4	3.87	6.15
2007	1.23	2.8	4.5	7.32
2008	1.67	3.67	5.79	9.07
2009	1.71	4.04	6.47	10.12
2010	2	4.64	7.49	11.69

Figure 18.4 shows the cyclical patterns of monthly data of 3-month interest rates of return on eurodollar deposits, US certificates of deposit (CDs), and Treasury bills during the period of 2000 to 2010.[1]

The National Bureau of Economic Research (NBER) and the US Department of Commerce have specified a number of time series as statistical business indicators of cyclical revivals and recessions. These time series have been classified into three groups.[2] The first group is the so-called *leading indicators,* such as the S&P index of the prices of 500 common stocks. These series have usually reached their cyclical turning points prior to the analogous turns in economic activity. The second group

[1] A eurodollar is any dollar on deposit outside the United States. In the bottom portion of Fig. 18.4, "spreads" are the differences between two different kinds of interest rates. For example, the eurodollar rate is 0.40 % higher than the US CD rate (0.27 %) and T-bill rate (0.14 %) in November 2010.

[2] Index numbers are essential elements for these business indicators. Therefore, we discuss these business indicators after we discuss index numbers and stock market indexes in the next chapter.

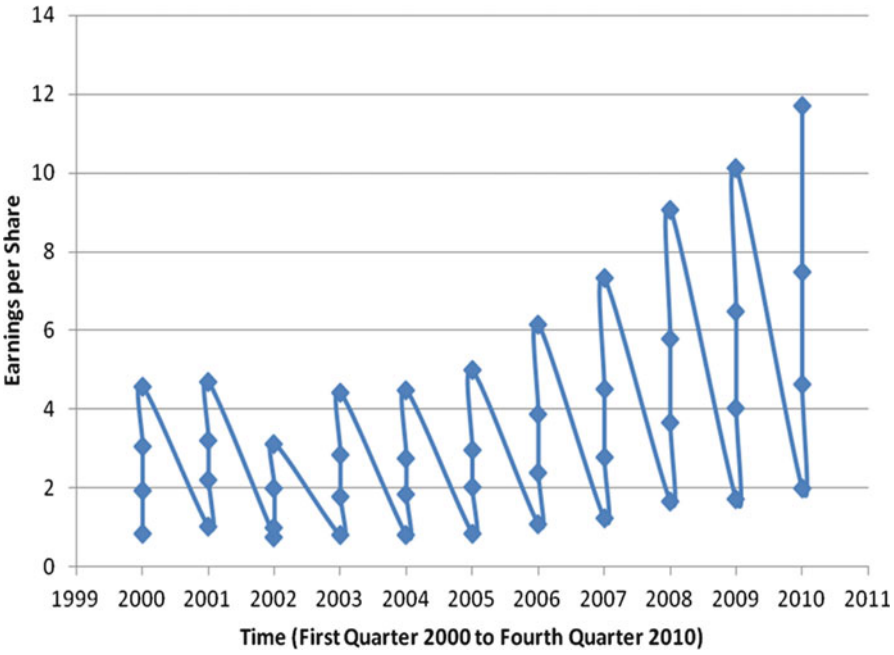


Fig. 18.2 Quarterly earnings per share of IBM

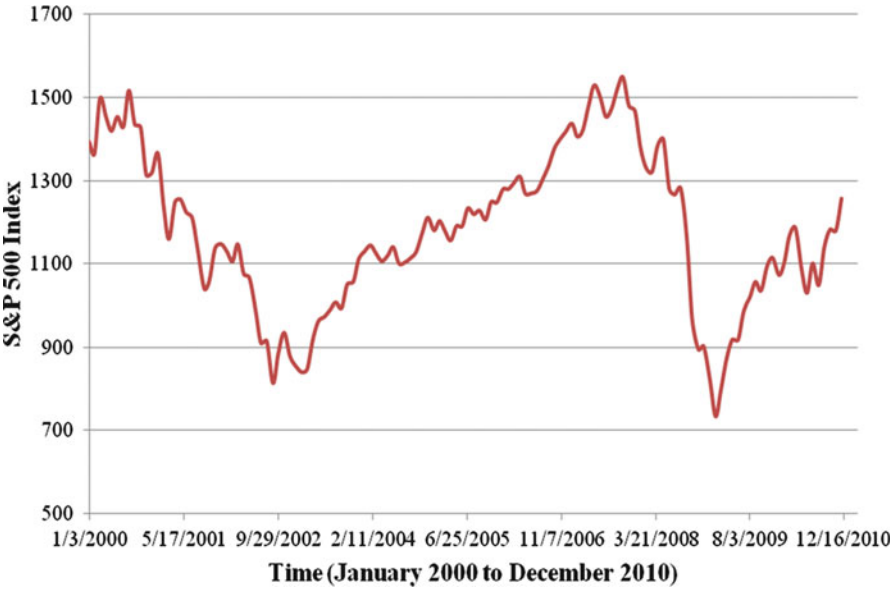


Fig. 18.3 S&P 500 Composite Index, January 2000 to December 2010

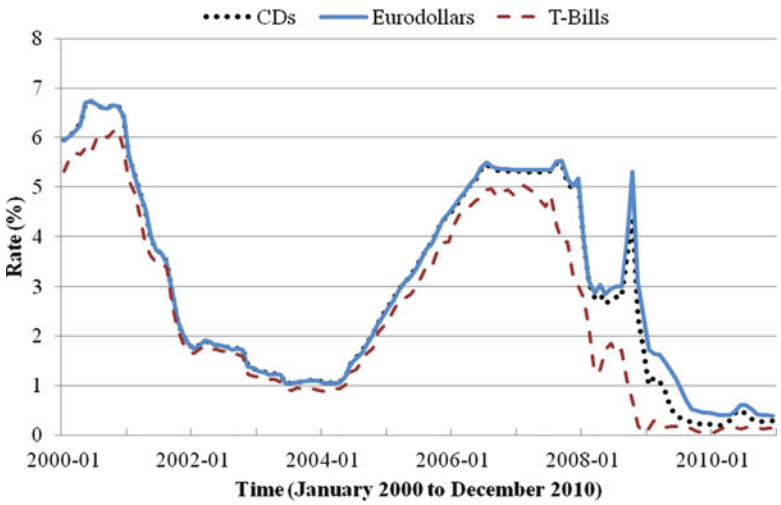


Fig. 18.4 Three-month rates on eurodollar deposits, US CDs, and US T-bills, 2000–2010 (monthly data)

is the *coincident indicators,* such as unemployment rate, the index of industrial production, and GNP in current dollars. The third group is the *lagging indicators,* such as index of labor cost per unit of output in manufacturing, business expenditures, and new plant and equipment. A particular indicator series is considered a leading, a coinciding, or a lagging indicator of overall economic activity, depending on whether the cyclical component of the series exhibits a tendency to precede, match, or follow the cyclical behavior of the economy at large.

18.2.4 The Irregular Component

The last component of the variation in a time series is the irregular element introduced by the unexpected event. For example, the announcement of a takeover bid may cause the price of the target company's stock to jump up 20 % or more in a single day. Fears of an outbreak of war in the Middle East and concerns about trade deficits and antitakeover legislation contributed to a spectacular decline in the stock market on October 19, 1987. And Iraq's invasion of Kuwait on August 3, 1990, caused worldwide stock markets to drop more than 10 % within a week. These irregular elements arise suddenly and have a temporary impact on time-series behavior.

Example 18.1 Graphical Presentation of Time-Series Components. In Fig. 18.5, Levenbach and Cleary show how a set of time-series data can be broken down into three components. Figure 18.5a is a plot of the original series of data. Figure 18.5b presents the trend component (long-term trend plus cyclical effects) of the series.

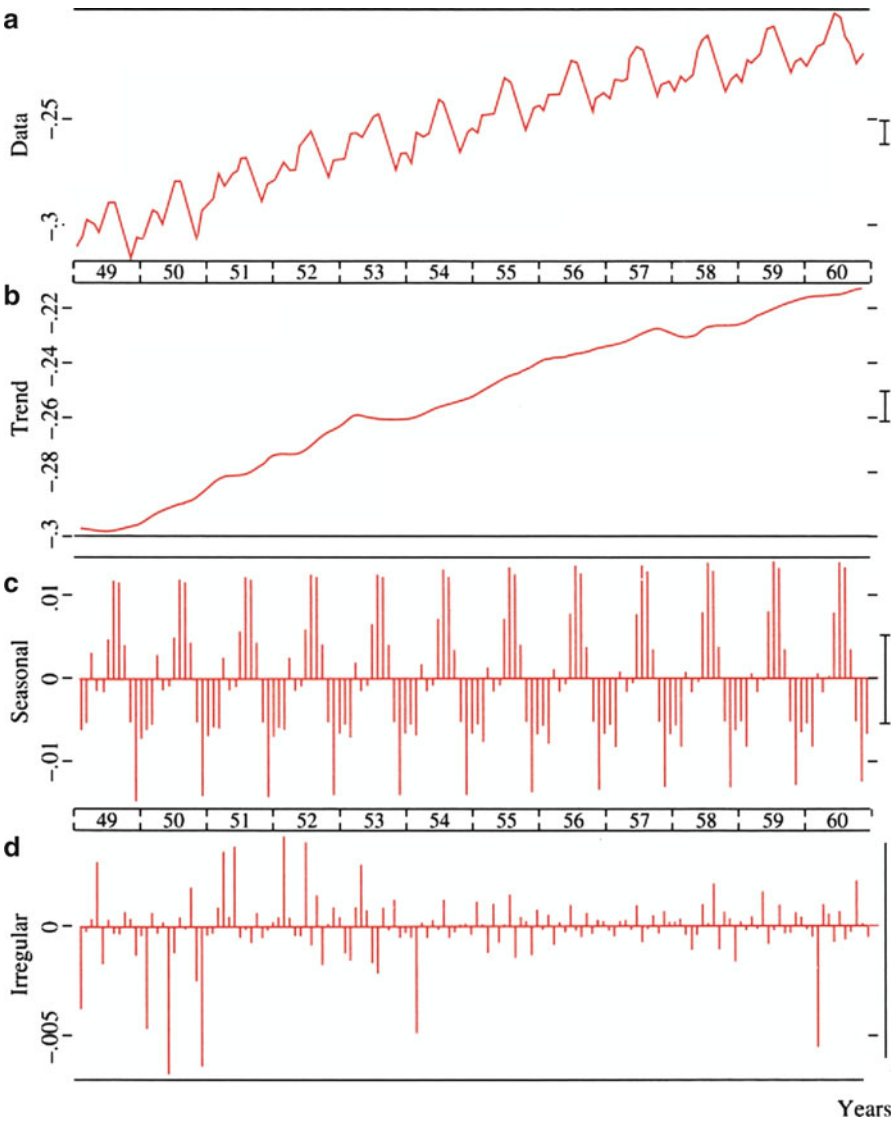


Fig. 18.5 Time-series decomposition (*Source:* Levenback, H., Cleary, J.P.: *The modem forecaster*, p. 50. Lifetime Learning Publications, New York (1984))

The data obviously exhibit an upward trend. Figure 18.5c presents the seasonal component of the data, and the irregular component appears in Fig. 18.5d.

Overall, a set of time-series data, x_t, can be described by using the additive model of Eq. 18.1 or the multiplicative model of Eq. 18.2:

$$x_t = T_t + C_t + S_t + I_t \tag{18.1}$$

$$x_t = T_t S_t C_t I_t \tag{18.2}$$

where

T_t = trend component
C_t = cyclical component
S_t = seasonal component
I_t = irregular component

For long-term planning and decision making in terms of time-series components, business executives concentrate primarily on forecasting the trend movement. For intermediate-term planning—say, from about 2 to about 5 years—fluctuations in the business cycle are of critical importance too. For short-term planning, and for purposes of operational decisions and control, seasonal variations must also be taken into account. In the next four sections of this chapter, we will analyze these time-series factors (components), and see how they can be forecasted.

18.3 Moving Average and Seasonally Adjusted Time Series

In this section, we explain how the moving-average method is used to smooth time-series data. We also discuss how moving average and related techniques can be used to obtain seasonally adjusted time-series data.

18.3.1 Moving Averages

Moving averages are usually associated with data smoothing. Smoothing a time series reduces the effects of seasonality and irregularity. As a result, the smoothed data reveal more information about seasonal trends and business cycles. The most common moving-average method is the unweighted moving average, in which each value of the data carries the same weight in the smoothing process. For a time series $x_1, \ldots, x_n$, the formula for doing a 3-term unweighted moving average is

$$z_t = \left(\frac{1}{3}\right) \sum_{i=0}^{2} x_{t-i} \quad (t = 3, \ldots, n) \tag{18.3}$$

Similarly, the k-term unweighted moving average is written

$$z_t = (1/k) \sum_{i=0}^{k-1} x_{t-i} \qquad (t = k, \ldots, n) \tag{18.4}$$

Alternatively, the weighted moving average can be used to replace the unweighted moving average. A k-term weighted moving average can be defined as

Table 18.3 Weighted average

(1) Observation value, x_{t-i}	(2) Weight, w_{t-i}	(3) $x_{t-i}w_{t-i}$
.035	.10	.0035
.002	.30	.0006
.100	.25	.0250
.060	*.35*	*.0210*
	1.00	*.0501* (weighted average)

$$z_t = \sum_{i=0}^{k-1} w_{t-i}x_{t-i} \qquad (t = k, \ldots, n) \tag{18.5}$$

where $\sum_{i=0}^{k-1} w_{t-i} = 1$

The w_{t-i}'s are known as weights and they sum to unity. If the w_{t-i}'s do not sum to unity, they can be normalized with a new set of weights (w^*_{t-i}) that sum to unity. The unweighted moving average is a special case of the weighted moving average with $w_i = 1/k$ for all i. An example of a weighted-average calculation appears in Table 18.3. Here, columns (1) and (2) represent observation value (x_{t-i}) and weight (w_i), respectively. Column (3) represents $x_{t-i}w_{t-i}$. From Table 18.3, we obtain

$$z_t = \sum_{i=0}^{3} x_{t-i}w_{t-i} = .0501$$

One of the important applications of moving averages is to deseasonalize seasonal time-series data which will be discussed in the next section.

18.3.2 Seasonal Index and Seasonally Adjusted Time Series

In Sect. 18.2, we noted that many business and economic time series contain a strong seasonal component. This component generally needs to be removed for either monthly or quarterly data. This section demonstrates how the moving-average procedure is used to remove the seasonal component and to do related analysis.

Suppose we have a quarterly time series, x_t, with a seasonal component. Then, we can apply Eq. 18.6, which is obtained by letting $k = 4$ in Eq. 18.4, to remove the seasonal component:

$$z_i = \left(\frac{1}{4}\right) \sum_{i=0}^{3} x_{t-i} \qquad (t = 4, \ldots, n) \tag{18.6}$$

Example 18.2 Seasonally Adjusted Quarterly Earnings per Share of Johnson & Johnson. For the data on quarterly earnings per share of J&J Corporation during the period of first quarter 2000 to fourth quarter 2010 given in Table 18.4, the first number in the series of the 4-quarter moving average is

$$\frac{.86 + 1.8 + 2.68 + 3.3}{4} = 2.16$$

and the second number is

$$\frac{1.8 + 2.68 + 3.3 + 1.0}{4} = 2.195$$

The complete series appears in column (3) of Table 18.4.

This 4-quarter moving-average time series is free from seasonally because it is always based on values such that each "season" is represented in each single observation of the new series (see Fig. 18.6). However, the location in time of the members of the series of moving averages does not correspond precisely with that of the members of the original series. Actually, the first 4-quarter moving average would be centered midway between the second-quarter and third-quarter dates. Hence, the 4-quarter moving-average series indicated in Eq. 18.6 should be rewritten either as

$$z_{t-.5} = \left(\frac{1}{4}\right)\sum_{i=2}^{-1} x_{t-i} \qquad (t = 3,\ 4,\ \ldots, n-2) \tag{18.7}$$

or

$$z_{t+.5} = \left(\frac{1}{4}\right)\sum_{i=-1}^{2} x_{t+i} \qquad (t = 2,\ 3,\ \ldots, n-2) \tag{18.7a}$$

Then the location-adjusted (centered) moving-average series can be written as

$$z_i^* = \frac{z_{t-.5} + z_{t+.5}}{2} \qquad (t = 3,\ 4,\ \ldots, n-2) \tag{18.8}$$

When

$$t = 3,\ z_3^* = \frac{z_{2.5} + z_{3.5}}{2} = \frac{x_1 + 2x_2 + 2x_3 + 2x_4 + x_5}{8}$$

The location-adjusted moving averages, z_t^*, are given in column (4) of Table 18.4. Both x_t and z_t^* are presented in Fig. 18.6.

We can use the location-adjusted moving-average data obtained from Eq. 18.8 to calculate seasonally adjusted series if we assume that the seasonal pattern through time is very stable. To do this, we need first to divide original data (x_t) by the location-adjusted moving averages (z_t^*) to obtain the *percentage of moving average*. That is,

Table 18.4 Actual (x_t) and centered 4-point moving average (z_t^*) earnings per share of Johnson & Johnson from first quarter 2000 to fourth quarter 2010

(1) t	(2) Earnings per share, x_t	(3) 4-point moving average, z_t	(4) Centered 4-point moving average, z_t^*
1	0.86		
2	1.8	2.16	
3	2.68	2.195	2.1775
4	3.3	1.995	2.095
5	1.0	1.7025	1.84875
6	1.0	1.345	1.52375
7	1.51	1.245	1.295
8	1.87	1.2825	1.26375
9	0.6	1.3375	1.31
10	1.15	1.42	1.37875
11	1.73	1.445	1.4325
12	2.2	1.43	1.4375
13	0.7	1.4475	1.43875
14	1.09	1.5025	1.475
15	1.8	1.5375	1.52
16	2.42	1.6825	1.61
17	0.84	1.8475	1.765
18	1.67	1.96	1.90375
19	2.46	1.99	1.975
20	2.87	2.03	2.01
21	0.96	2.085	2.0575
22	1.83	2.2125	2.14875
23	2.68	2.25	2.23125
24	3.38	2.31	2.28
25	1.11	2.3925	2.35125
26	2.07	2.4875	2.44
27	3.01	2.4325	2.46
28	3.76	2.4025	2.4175
29	0.89	2.36	2.38125
30	1.95	2.3375	2.34875
31	2.84	2.4325	2.385
32	3.67	2.5575	2.495
33	1.27	2.7575	2.6575
34	2.45	2.995	2.87625
35	3.64	2.995	2.995
36	4.62	2.99	2.9925
37	1.27	2.99	2.99
38	2.43	2.9475	2.96875
39	3.64	3.04	2.99375
40	4.45	3.155	3.0975
41	1.64	3.28	3.2175
42	2.89	3.38	3.33
43	4.14		
44	4.85		

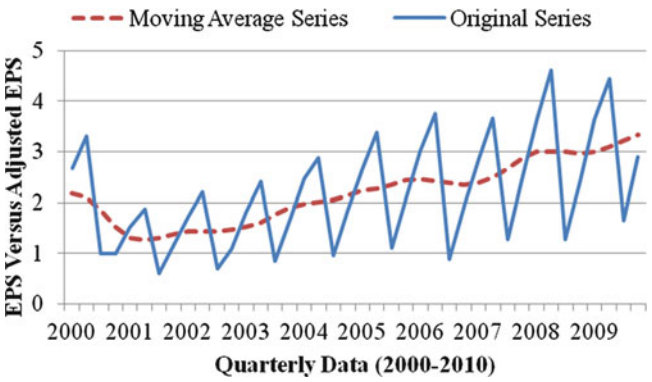


Fig. 18.6 Earnings per share versus moving-_average EPS for Johnson & Johnson

$$\text{Percentage of moving average (PMA)} = 100\left(\frac{x_t}{z_t^*}\right) \tag{18.9}$$

The PMA of earnings per share for Johnson & Johnson is presented in column (4) of Table 18.5.

In our case, the first observation of PMA is

$$100\left(\frac{x_3}{z_3^*}\right) = 100\left(\frac{2.68}{2.1775}\right) = 123.0769$$

We assume that for any given quarter, in each year, the effect of seasonality is to raise or lower the observation by a constant proportionate amount (*seasonal index*) compared with what it would have been in the absence of seasonal influences. Then we use the so-called seasonal index method lo remove the seasonal component.

Let's explore the logic of and procedure for calculating the seasonal index listed in column (5) of Table 18.5. By dividing z_t^* into x_t, we can explicitly write the percentage of moving average as

$$100\left(\frac{x_t}{z_t^*}\right) = \frac{100T_tC_tS_tI_t}{T_tC_t} = 100S_tI_t \tag{18.10}$$

The $100S_tI_t$ series for earnings per share of Johnson & Johnson is presented in Fig. 18.7. This series contains both seasonal and irregular components. The next step is to remove the effect of irregular movements from $100(x_t/z_t^*)$. We do this by taking the median of the percentage of moving-average figures for the same quarter as indicated in Table 18.6. The medians for the first through the fourth quarters are 47.400, 84.122, 121.152, and 149.278, respectively. The total of these medians is 401.953. It is desirable that the total of the 4 indexes be 400, in order that they

Table 18.5 Seasonal adjustment of earnings per share of Johnson & Johnson by the Seasonal Index Method from first quarter 2000 to fourth quarter 2010

(1) Date	(2) EPS, x_t	(3) z_t^*	(4) $100(x_t/z_t^*)$	(5) Seasonal index	(6) Adjusted EPS [Col. (2) ÷ Col. (5)] × 100
2000.1	0.86			47.1702	1.823185
2	1.8			83.71375	2.150184
3	2.68	2.1775	123.0769	120.5633	2.2229
4	3.3	2.095	157.5179	148.5528	2.221432
2001.1	1	1.84875	54.0906	47.1702	2.119983
2	1	1.52375	65.62756	83.71375	1.194547
3	1.51	1.295	116.6023	120.5633	1.252455
4	1.87	1.26375	147.9723	148.5528	1.258812
2002.1	0.6	1.31	45.80153	47.1702	1.27199
2	1.15	1.37875	83.40888	83.71375	1.373729
3	1.73	1.4325	120.7679	120.5633	1.434931
4	2.2	1.4375	153.0435	148.5528	1.480955
2003.1	0.7	1.43875	48.65334	47.1702	1.483988
2	1.09	1.475	73.89831	83.71375	1.302056
3	1.8	1.52	118.4211	120.5633	1.492992
4	2.42	1.61	150.3106	148.5528	1.62905
2004.1	0.84	1.765	47.59207	47.1702	1.780785
2	1.67	1.90375	87.7216	83.71375	1.994893
3	2.46	1.975	124.557	120.5633	2.040423
4	2.87	2.01	142.7861	148.5528	1.931973
2005.1	0.96	2.0575	46.65857	47.1702	2.035183
2	1.83	2.14875	85.16579	83.71375	2.186021
3	2.68	2.23125	120.112	120.5633	2.2229
4	3.38	2.28	148.2456	148.5528	2.275285
2006.1	1.11	2.35125	47.20893	47.1702	2.353181
2	2.07	2.44	84.83607	83.71375	2.472712
3	3.01	2.46	122.3577	120.5633	2.496615
4	3.76	2.4175	155.5326	148.5528	2.531087
2007.1	0.89	2.38125	37.37533	47.1702	1.886785
2	1.95	2.34875	83.02288	83.71375	2.329366
3	2.84	2.385	119.0776	120.5633	2.35561
4	3.67	2.495	147.0942	148.5528	2.470502
2008.1	1.27	2.6575	47.78928	47.1702	2.692378
2	2.45	2.87625	85.18036	83.71375	2.92664
3	3.64	2.995	121.5359	120.5633	3.019162
4	4.62	2.9925	154.386	148.5528	3.110005
2009.1	1.27	2.99	42.47492	47.1702	2.692378
2	2.43	2.96875	81.85263	83.71375	2.902749
3	3.64	2.99375	121.5866	120.5633	3.019162
4	4.45	3.0975	143.6642	148.5528	2.995568
2010.1	1.64	3.2175	50.97125	47.1702	3.476772
2	2.89	3.33	86.78679	83.71375	3.45224
3	4.14			120.5633	3.433882
4	4.85			148.5528	3.264833

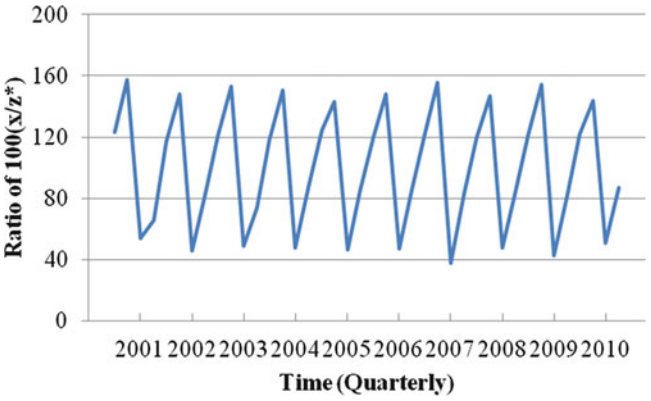


Fig. 18.7 Trend of $100(x_t/z_t^*)$ ratio for Johnson & Johnson

Table 18.6 Calculation of seasonal indexes of EPS for Johnson & Johnson Corporation

	Quarter				
Year	*1*	*2*	*3*	*4*	Sums
2000			123.077	157.518	
2001	54.091	65.628	116.602	147.972	
2002	45.802	83.409	120.768	153.043	
2003	48.653	73.898	118.421	150.311	
2004	47.592	87.722	124.557	142.786	
2005	46.659	85.166	120.112	148.246	
2006	47.209	84.836	122.358	155.533	
2007	37.375	83.023	119.078	147.094	
2008	47.789	85.180	121.536	154.386	
2009	42.475	81.853	121.587	143.664	
2010	50.971	86.787			
Median	47.400	84.122	121.152	149.278	401.953
Seasonal index	47.170	83.714	120.563	148.553	400.000

average 100 %, so we multiply each of them by an adjustment factor (400/401.953) to make the sum of the 4-quarter seasonal indexes equal 400. The seasonal index is presented in column (5) of Table 18.5.[3] Dividing the seasonal index into the original quarterly data and multiplying the result by 100, we obtain the adjusted series presented in column (6) of Table 18.5 and in Fig. 18.8.

This seasonal index method of seasonal adjustment shows us one possible and simple way to solve the problem of eliminating the seasonal component. In practice, however, it generally can be solved by computer. Important government monthly and quarterly economic data such as consumer price indexes and employment and unemployment rates have strong seasonal components, and government agencies generally publish these data in both unadjusted and adjusted forms. The seasonal

[3] The mean instead of the median can also be used to calculate the seasonal index.

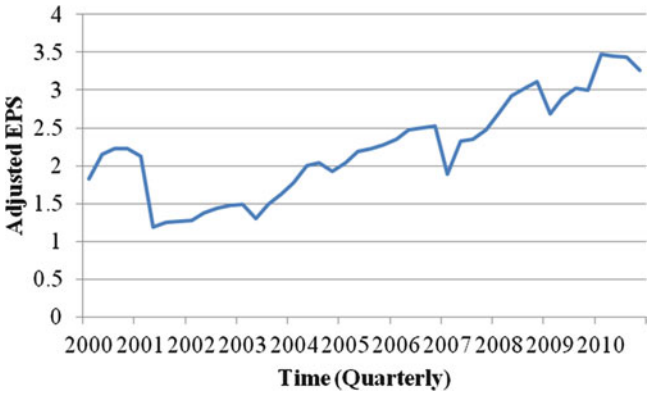


Fig. 18.8 Adjusted Earnings per Share (EPS) of Johnson & Johnson

adjustment procedure used in official United States government publications is the Census X-11 method which is based upon the moving-average method.[4] In the next section, we will look at time trend regression.

18.4 Linear and Log-Linear Time Trend Regressions

If a time series is expected to change linearly overtime, the simple linear regression model defined in Eq. 18.11 can be used to relate the time series, x_t, to time t, and the least-squares line is used to forecast future values of x_t:

$$x_t = \alpha + \beta t + \varepsilon_t \tag{18.11}$$

If the relationship between x_t and t is multiplicative instead of additive, then transforming x_t by taking the natural logarithm enables us to make the relationship linear. For example, let x_0 and x_t be the sales of a firm in the base year and in year t, respectively. Then the underlying relationship is

$$x_t = x_0 e^{gt}$$

where x_0 is the base-year sales figure, g is the growth rate, and t is the length of time in terms of number of periods. Then, via the natural logarithm transformation, we obtain

$$\begin{aligned} \log_e x_t &= \log_e (x_0 e^{gt}) \\ &= \log_e x_0 + gt \end{aligned} \tag{18.12}$$

[4] The X-11 model for decomposing time series components can be found in Appendix 24.A of the book entitled "Financial Analysis, Planning and Forecasting: Theory and Application" by Lee, A. C, J. C. Lee and C. F. Lee., 2nd ed. Singapore: World Scientific Publishing Company, 2009.

where $\log_e$ is the natural logarithm operator. Equation 18.12 can be defined as a log-linear regression model[5]:

$$\log_e x_t = \alpha' + \beta' t + \varepsilon_t' \tag{18.13}$$

where

$\alpha' = \log_e x_0$
$\beta' = g =$ growth rate of a firm's sales

JNJ's annual sales data (1980–2010), presented in Table 18.7, are used to show how Eq. 18.12 can be employed to forecast JNJ's future sales, and Eq. 18.13 to estimate the growth rate of JNJ's historical sales.

Example 18.3 Forecasting Sales and Estimating Growth Rate. Suppose Johnson & Johnson Company is interested in forecasting its sales revenues for each of the next 6 years. The sales manager of the company would also like to estimate the historical growth rate of sales revenue.

To make forecasts and assess their reliability, we must construct a time-series model for the sales revenue data listed in Table 18.7. A plot of the data (Fig. 18.9) reveals a linearly increasing trend. Therefore, the linear time trend regression defined in Eq. 18.11 can be used to do forecasting. By the method of least squares (see Sect. 13.3), we obtain the least-squares model in terms of sales (x_t) and time intervals (t) as

$$\hat{x}_t = \hat{\alpha} + \hat{\beta} t = -7965.026 + 2089.257t$$

With $R^2 = .903$.

This least-squares line is shown in Fig. 18.9, and the result of straight-line model is given in Fig. 18.10. We can now forecast sales for years 2011–2016 by log-linear regression model defined in Eq. 18.13. The forecasts of sales by log-linear regression model and the corresponding 95 % prediction intervals are shown in Fig. 18.11. Although it is not easily perceptible in the figure, the prediction interval widens as we attempt to forecast further into the future. This agrees with the intuitive notion that short-term forecasts should be more reliable than long-term forecasts.

To estimate the growth rate for JNJ's sales during the period 1980–2010, we use data listed in Table 18.7 to fit the log-linear regression of Eq. 18.13 and obtain

$$\log_e \hat{x}_t = \hat{\alpha}' + \hat{\beta}' t = 8.3005 + .0944t$$
$$(0.027) \quad (0.001) \quad R^2 = .993$$

[5] In this regression, we implicitly assume that x_t is lognormally distributed and that $\log_e x_t$ is normally distributed. The relationship between the normal and lognormal distributions was discussed in Sect. 7.4 of Chap. 7.

Table 18.7 Johnson & Johnson's annual sales

Year	Sales, x_t (in millions)	t
1980	$4,837.38	1
1981	$5,399.00	2
1982	$5,760.87	3
1983	$5,972.87	4
1984	$6,124.50	5
1985	$6,421.30	6
1986	$7,002.90	7
1987	$8,012.00	8
1988	$9,000.00	9
1989	$9,757.00	10
1990	$11,232.00	11
1991	$12,447.00	12
1992	$13,753.00	13
1993	$14,138.00	14
1994	$15,734.00	15
1995	$18,842.00	16
1996	$21,620.00	17
1997	$22,629.00	18
1998	$23,657.00	19
1999	$27,471.00	20
2000	$29,139.00	21
2001	$33,004.00	22
2002	$36,298.00	23
2003	$41,862.00	24
2004	$47,348.00	25
2005	$50,434.00	26
2006	$53,194.00	27
2007	$61,035.00	28
2008	$63,747.00	29
2009	$61,897.00	30
2010	$61,587.00	31

Figures in parentheses are standard errors. This result implies that the estimated growth rate $g = \hat{\beta}' = 9.44\ \%$. In other words, the annual growth rate of JNJ's sales was 9.44 % during the period 1980–2010.

18.5 Exponential Smoothing and Forecasting

18.5.1 Simple Exponential Smoothing and Forecasting

Smoothing techniques are often used to forecast future values of a time series. One problem that arises in using a moving average to forecast time series is that values at the ends of the series are lost, as shown in Sect. 18.3. Therefore, we must

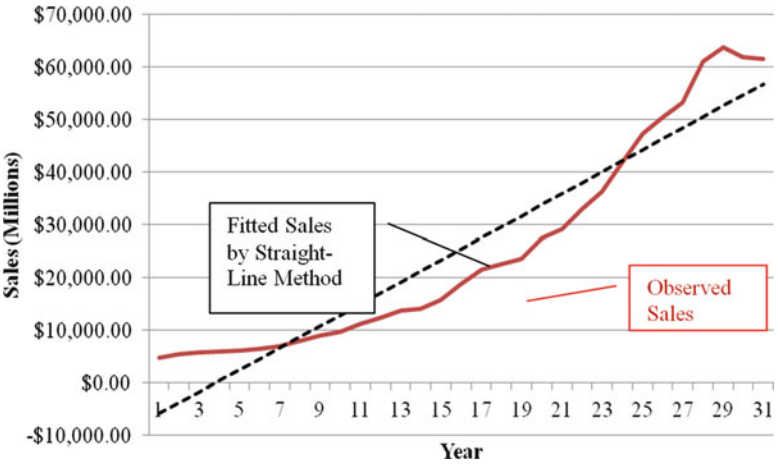


Fig. 18.9 J&J's annual sales (1980–2010) and the linear trend regression

Summary Output

Regression statistics	
Multiple R	0.950
R square	0.903
Adjusted R square	0.899
Standard error	6346.179
Observations	31.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.000	10825189941.84	10825189941.84	268.789	0.000
Residual	29.000	1167945584.055	40273985.65		
Total	30.000	11993135525.89			

	Coefficients	*Standard error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7965.026	2335.910	-3.410	0.002	-12742.498	-3187.553
t	2089.257	127.434	16.395	0.000	1828.625	2349.890

Fig. 18.10 Least-squares fit (straight-line method) to x_t = Sales

subjectively extend the graph of the moving average into the future. No exact calculation of a forecast is available, because generating the moving average at a future time period t requires that we know one or more future values of the series. A technique that leads to forecasts that can be explicitly calculated is called *exponential smoothing*. To use the exponential smoothing technique in forecasting, we need only past and current values of the time series.

To obtain an exponentially smoothed series, we first need to choose a weight α between 0 and 1 called the *exponential smoothing constant*. The exponentially smoothed series, denoted s_t, is then calculated as follows:

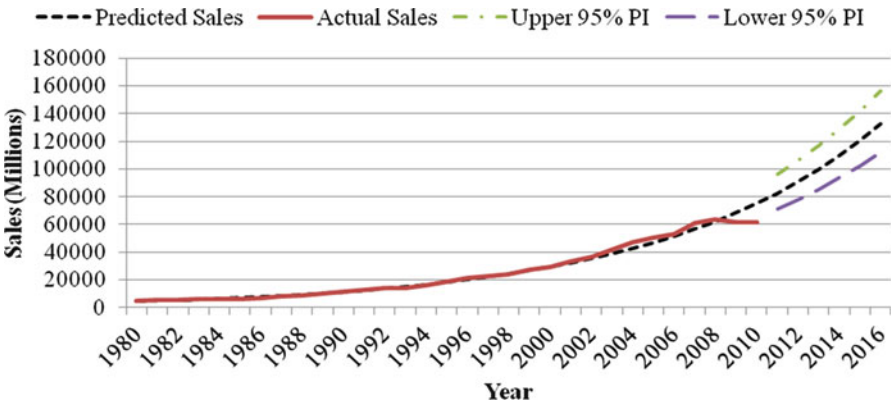


Fig. 18.11 Observed (years 1980–2010) and forecast (years 2011–2016) sales using log-linear regression model

$$\begin{aligned} s_1 &= x_1 \\ s_2 &= \alpha x_2 + (1-\alpha)s_1 \\ s_3 &= \alpha x_3 + (1-\alpha)s_2 \\ &\vdots \qquad\qquad \vdots \\ s_i &= \alpha x_t + (1-\alpha)s_{t-1} \end{aligned} \tag{18.14}$$

We can see that the exponentially smoothed value at time t is simply a weighted average of the current time-series value x_t and the exponentially smoothed value at the previous time period, S_{t-1}. Then we can use s_t to do forecasting as follows:

$$\hat{x}_{t+1} = s_t = \alpha x_t + (1-\alpha)s_{t-1} \tag{18.15}$$

where $\hat{x}_{t+1}$ is the next period's forecast value. In other words, $\hat{x}_{t+1}$ is expressed in terms of the smoothing constant times x_t plus (1–α) times s_{t-1}.

If the manager of a company in 1990 ($t = 1$) knows only that current sales of his or her company equal $x_1 = 5{,}000$ units and that current sales have been forecasted as $s_0 = 5{,}100$ units, then he or she can use Eq. 18.15 to forecast 1991 sales. If we choose $\alpha = .30$ as a smoothing constant, then the sales for 1991 are forecasted in terms of Eq. 18.15 as

$$\hat{x}_2 = s_1 = (.30)(5,000) + (1-.30)(5,100) = 5,070 \text{units}$$

Rewriting Eq. 18.15 as

$$\hat{x}_{t+1} = s_t = s_{t-1} + \alpha(x_t - s_{t-1}) \tag{18.16}$$

implies that simple exponential smoothing is the weighted average of s_{t-1} and the forecast error ($x_t - s_{t-1}$) with weights of 1 and α, respectively. The term *exponential*

smoothing refers to the fact that s_t can be expressed as a weighted average with exponentially decreasing weights, as we now illustrate.

We substitute the expressions for s_{t-1} and s_{t-2} into the expression for s_t as denned in Eq. 18.15 and obtain

$$s_{t-1} = \alpha x_{t-1} + (1-\alpha)s_{t-2}$$
$$s_{t-2} = \alpha x_{t-2} + (1-\alpha)s_{t-3}$$

Repeatedly substituting s_{t-2} and s_{t-1} into Eq. 18.15 reveals that

$$\begin{aligned} s_t &= \alpha x_t + (1-\alpha)s_{t-1} \\ &= \alpha x_t + \alpha(1-\alpha)x_{t-1} + (1+\alpha)^2 s_{t-2} \\ &= \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + (1-\alpha)^3 s_{t-3} \end{aligned}$$

Continuous substitution for $s_{t-k,}$ where $k = 2, 3,\ldots, t$, yields

$$s_t = \left[\alpha \sum_{k=0}^{t-1} (1-\alpha)^k x_{t-k}\right] + (1-\alpha)^t s_0 \qquad (0<\alpha<1) \qquad (18.17)$$

where s_0 is an initial estimate of the smoothed value.

The sum of weights approaches unity as t approaches infinity; hence, we use the term *average*.[6] The weights decrease geometrically with increasing k, so the most recent values of x_t are assigned the greatest weight. Ail the previous values of x_t are included in the expression for s_t. Because α is less than unity, the most remote values *of* x_t are associated with the smallest weights. The selection of α depends on the sensitivity of the response required by the model. For example, a small α is used to represent the small sensitivity of the response, and it implies that a single change won't affect the moving average much. The smaller the value of α, the slower the response. Note that the method discussed in this section is good only for short-term forecasting.

In the next example, we draw on annual earnings per share (EPS) data for both Johnson & Johnson (J&J) and International Business Machines (IBM) to show how the simple exponential smoothing method defined in Eq. 18.15 can be used to do data analysis.

[6] Let $0 < \alpha \leq 1$, as $t \geq \infty$, $(1-\alpha)^t \geq 0$. Let

$$y = \alpha + \alpha(1-\alpha) + \alpha(1-\alpha)^2 + \cdots + \alpha(1-\alpha)^{t-1} \qquad (A)$$
$$(1-\alpha)y = \alpha(1-\alpha) + \alpha(1-\alpha)^2 + \cdots + \alpha(1-\alpha)^t + \cdots \qquad (B)$$

Subtracting Equation B from Equation A yields $y = 1-(1-\alpha)^t$. Because $\alpha < 1$, y approaches 1 if t approaches infinity. This implies that $\alpha + \alpha(1-\alpha) + \alpha(1-\alpha)^2 + \ldots = 1$.

Example 18.4 Simple Exponential Smoothing of EPS for Both J&J and IBM. Consider the EPS for both J&J and IBM from 2000 to 2010 as shown in the second column of Table 18.8. Using $\alpha = .3$, we calculate the exponentially smoothed series presented in the third column of Table 18.8 as follows:

IBM	JNJ
$s_{00} = x_{00} = 4.58$	$s_{00} = x_{00} = 3.45$
$s_{01} = .3(4.45) + .7(4.58) = 4.541$	$s_{01} = .3(1.87) + .7(3.45) = 2.976$
$\vdots$	$\vdots$
$s_{10} = .3(11.69) + .7(7.734566)$ $= 8.921196$	$s_{10} = .3(4.85) + .7(3.939735)$ $= 4.212814$

We see from the table that the most recent estimates of smoothed EPS for J&J and IBM are

$$s_n = s_{10} = 8.921196 \quad \text{(IBM)}$$
$$s_n = s_{10} = 4.212814 \quad \text{(J\&J)}$$

These values are then used as the forecast of EPS for both J&J and IBM for future years. The observed series and these forecasts for J&J and IBM are graphed in Figs. 18.12 and 18.13, respectively.

Finally, note that the choice of the smoothing constant (α) affects the precision of the forecast. In practice, we can try several different values to see which would have been most successful in predicting historical movement in the time series. For example, we might compute the smoothed series for values of α of .3, .4, .5, and .7 and calculate the forecast *mean squared error* (*MSE*) for these four different α-values:

$$\text{MSE} = \frac{\sum_{t=1}^{n} (x_t - \hat{x}_t)^2}{n} \tag{18.18}$$

where x_t and $\hat{x}_t$ are actual value and forecast value, respectively. The value of α for which this MSE is smallest is then used in the prediction of future values.

18.5.2 The Holt–Winters Forecasting Model for Nonseasonal Series

The simple exponential smoothing technique discussed in the previous section does not recognize the trend in the time series. In this section, we will generalize the

Table 18.8 Simple exponential smoothing ($\alpha = .3$) of EPS for J&J and lBM

t	x_t	s_t
IBM		
2000	4.58	4.58
2001	4.45	4.541
2002	2.1	3.8087
2003	4.4	3.98609
2004	5.03	4.299263
2005	4.96	4.497484
2006	6.2	5.008239
2007	7.32	5.701767
2008	9.07	6.712237
2009	10.12	7.734566
2010	11.69	8.921196
J&J		
2000	3.45	3.45
2001	1.87	2.976
2002	2.2	2.7432
2003	2.42	2.64624
2004	2.87	2.713368
2005	3.5	2.949358
2006	3.76	3.19255
2007	3.67	3.335785
2008	4.62	3.72105
2009	4.45	3.939735
2010	4.85	4.212814

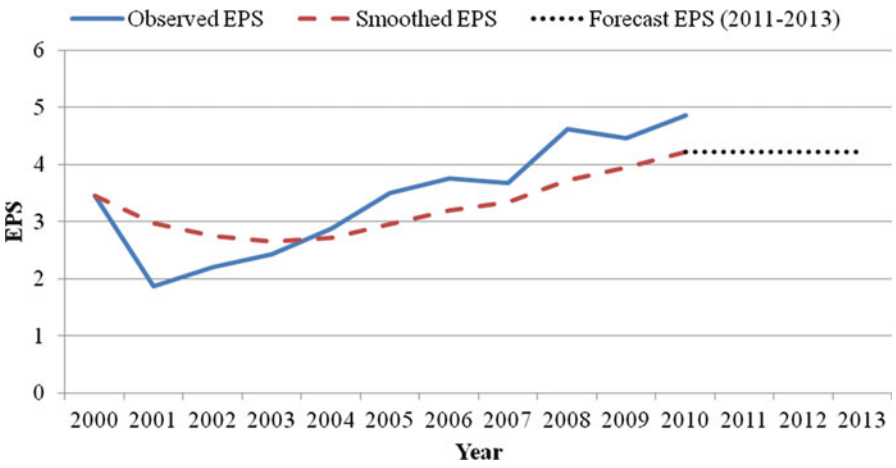


Fig. 18.12 Annual earnings per share of J&J (simple exponential smoothing)

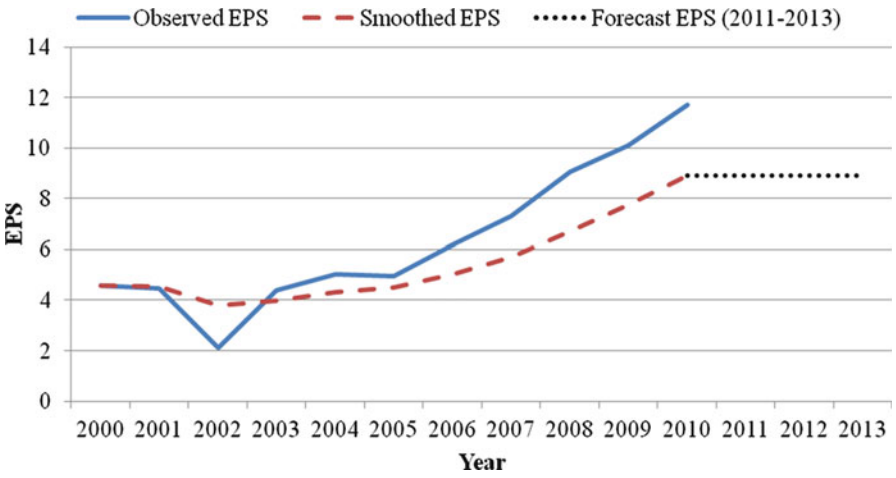


Fig. 18.13 Annual earnings per share of IBM (simple exponential smoothing)

simple exponential smoothing model defined in Eq. 18.15 by explicitly recognizing the trend in a time series. The *Holt–Winters forecasting model*[7] consists of both an exponentially smoothed component (s_t) and a trend component (T_t). The trend component is used in calculating the exponentially smoothed value. Here, s_t and T_t can be written as

$$s_t = \alpha x_t + (1 - \alpha)(s_{t-1} + T_{t-1}) \tag{18.19a}$$

$$T_t = \beta(s_t - s_{t-1}) + (1 - \beta)T_{t-1} \tag{18.19b}$$

where α and β are two smoothing constants, each of which is between 0 and 1. We estimate the trend component of the series by using a weighted average of the most recent change in the smoothed component [represented by $(s_t - s_{t-1})$] and the time trend estimate (represented by T_{t-1}) from the previous period. The procedure for calculating the Holt–Winters components is as follows:

1. Choose an exponential smoothing constant a between 0 and 1. Small values of α give less weight to the current values of the time series and more weight to the past. Large values of α give more weight to the current values of the series.
2. Choose a trend smoothing constant β between 0 and 1. Small values of β give less weight to the current changes in the level of the series and more weight to the past trend. Larger choices assign more weight to the most recent trend of the series.
3. Estimate the first observation of trend T_1 by one of the following two alternative methods.

[7] The Holt–Winters forecasting model for seasonal series will be discussed in Appendix 1 of Chap. 18.

Method I:
Let $T_1 = 0$. If there are a large number of observations in the time series, this method provides an adequate initial estimate for the trend.
Method 2:
Use the first 5 (or so) observations to estimate the initial trend by following the linear time trend regression line

$$x_t = a + bt + e_t$$

Then use the estimated slope $\hat{b}$ as the first trend observation; that is, $T_1, = \hat{b}$.

4. Calculate the components s_t and T_1 from the time series as follows:

$$\begin{aligned} s_1 &= x_1 \\ T_1 &= 0 \text{ or } \hat{b} \\ s_2 &= \alpha x_2 + (1-\alpha)(s_1 + T_1) \\ T_2 &= \beta(s_2 - s_1) + (1-\beta)T_t \\ &\vdots \\ s_t &= \alpha x_t + (1-\alpha)(s_{t-1} + T_{t-1}) \\ T_t &= \beta(s_t - s_{t-1}) + (1-\beta)T_{t-1} \end{aligned}$$

The data on earnings per share of J&J and IBM listed in Table 18.8 show how the forecasting model defined in Eqs. 18.19a and 18.19b can be used to do data analysis.

Example 18.5 Using the Holt–Winters Model to Estimate the EPS of J&J and IBM. Now let's use the Holt–Winters model to do the exponential smoothing for the EPS data for both J&J and IBM listed in Table 18.9. We begin by using the first five observations to estimate the first term of the trend component. The estimated slopes for the EPS of J&J and IBM are 0 and 1.275, respectively. Let $\alpha = .3$ and $\beta = .2$. Following the formula for the Holt–Winters components listed in step 4, we calculate

J&J	**IBM**
$s_1 = x_1 = 3.45$	$s_1 = x_1 = 4.58$
$T_1 = 0$	$zT_1 = 0.085$
$s_2 = .3(1.87) + .7(3.45 + 0)$	$s_2 = .3(4.45) + .7(4.58 + 0.085)$
$= 2.976$	$= 4.6005$
$T_2 = .2(2.976 - 3.45) + .8(0)$	$T_2 = .2(4.6005 - 4.58) + .8(0.085)$
$= -0.0948$	$= 0.0721$
$\vdots$	$\vdots$

The remaining calculations are carried out in precisely the same way. All s_t- and T_1-values for both J&J and IBM are given in Table 18.9.

How are these estimates of EPS level and trend used to forecast future observations? Given a series $x_1, x_2, \ldots, x_n$, the most recent EPS level and trend estimates are

Table 18.9 EPS for IBM and J&J and their smoothed series in terms of the Holt–Winters forecasting model

t	x_t	s_t	T_t
IBM			
2000	4.58	4.58	0.085
2001	4.45	4.6005	0.0721
2002	2.1	3.90082	−0.08226
2003	4.4	3.992995	−0.04737
2004	5.03	4.270937	0.017693
2005	4.96	4.490041	0.057975
2006	6.2	5.043611	0.157094
2007	7.32	5.836494	0.284252
2008	9.07	7.005522	0.461207
2009	10.12	8.26271	0.620403
2010	11.69	9.725179	0.788816
J&J			
2000	3.45	3.45	0
2001	1.87	2.976	−0.0948
2002	2.2	2.67684	−0.13567
2003	2.42	2.504818	−0.14294
2004	2.87	2.514313	−0.11245
2005	3.5	2.731301	−0.04657
2006	3.76	3.007314	0.01795
2007	3.67	3.218685	0.056634
2008	4.62	3.678723	0.137315
2009	4.45	4.006227	0.175353
2010	4.85	4.382105	0.215458

s_n and T_n, respectively. To do forecasting, we assume that the latest trend will continue from the most recent level. In general, standing at time n and looking m time periods into the future, we define the prediction for the m period ahead as

$$\hat{x}_{t+m} = s_t + mT_t \tag{18.20}$$

If $T_t = 0$, then this prediction reduces to the simple exponential smoothing prediction discussed in Example 18.3. On the basis of this formula and the information given in Table 18.9, we calculate the future predictions for both J&J and IBM as

J&J	IBM
$s_{2011} = 4.382105 + .215458 = 4.597563$	$s_{2011} = 9.725179 + .788816 = 10.514$
$s_{2012} = 4.382105 + (2)(.215458)$	$s_{2012} = 9.725179 + (2)(.788816)$
$= 4.813021$	$= 11.30281$
$s_{2013} = 4.382105 + (3)(.215458)$	$s_{2013} = 9.725179 + (3)(.788816)$
$= 5.028479$	$= 12.09163$

Figures 18.14 and 18.15 show the data series and three forecasts for J&J and IBM, respectively.

Finally, note that the choice of smoothing constants (α and β) affects the precision of a forecast. In practice, we can try several different values of α and β to see which would have been most successful in predicting historical movement in

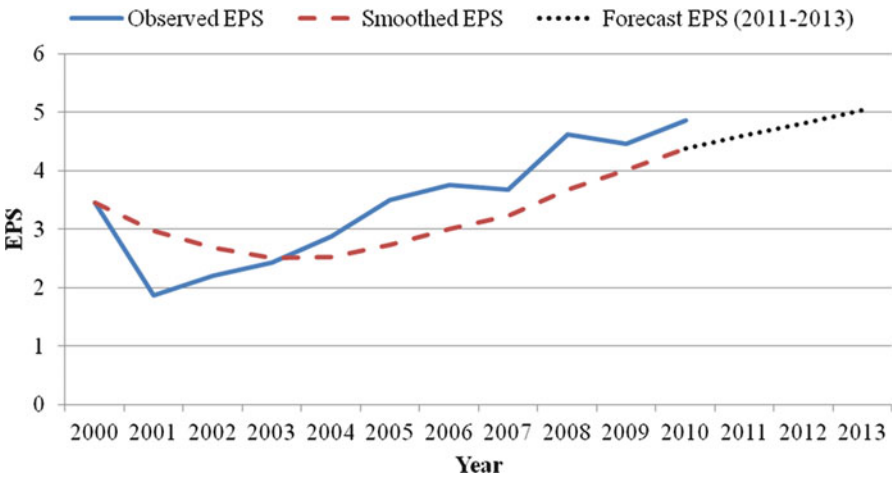


Fig. 18.14 Annual earnings per share of J&J with forecasts based on the Holt–Winters Model

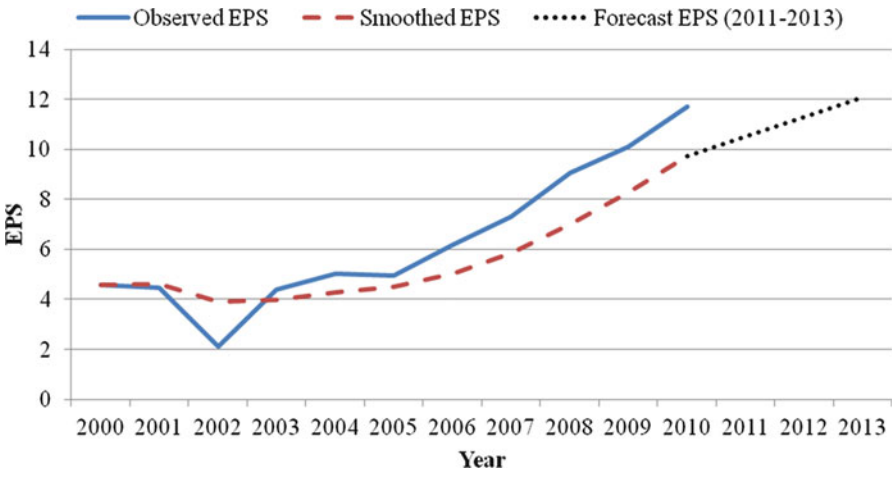


Fig. 18.15 Annual earnings per share of lBM with forecasts based on the Holt–Winters Model

the time series. Again, the forecast mean squared error as defined in Eq. 18.18 can be used as a benchmark in deciding what values of α and β are appropriate for forecasting future observations.

18.6 Autoregressive Forecasting Model

A time-series analysis always reveals some degree of correlation between elements. For example, a certain firm's current sales may be correlated with sales in the previous period and even with sales in several prior periods. Under these

circumstances, we can regress the time series x_t on some combination of its past values to derive a forecasting equation.

Suppose we attempt to predict the value of x_t by using previous observation. The prediction equation is

$$\hat{x}_t = a_0 + a_1 x_{t-1} \tag{18.21}$$

where α_0 and α_1 are the least-squares regression estimates. This is called a first-order *autoregressive forecasting model*[8], AR(1). If the current value of a time series depends on the two most recent observations, we can use the model

$$\hat{x}_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} \tag{18.22}$$

where a_0, a_1, and a_2 are least-squares regression estimates. This is called a second-order autoregressive model, AR(2). Generally, the autoregressive model of order p, AR(P), can be expressed as

$$\hat{x}_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} \tag{18.23}$$

where a_0, a_1, a_2,. . ., a_p are least-squares regression estimates.

In the next example, quarterly data on Johnson & Johnson's sales are employed to show how the autoregressive model can be used in forecasting.

Example 18.6 Sales Forecast for Johnson & Johnson. Quarterly sales data for Johnson & Johnson from first quarter 2000 through fourth quarter 2010 are presented in Table 18.10 and Fig. 18.16.

Using the data in Table 18.10, we run the AR(1), AR(2), and AR(3) models:

$$\begin{aligned} \text{AR}(1): \text{Sales}_t &= 552.7913 + 0.9703\ \underset{(0.026)}{\text{sales}_{t-1}} \\ R^2 &= .9719 \end{aligned} \tag{18.24}$$

$$\begin{aligned} \text{AR}(2): \text{Sales}_t &= 586.6586 + \underset{(.1623)}{.9106\ \text{sales}_{t-1}} + \underset{(.1590)}{.0580\ \text{sales}_{t-2}} \\ R^2 &= .9702 \end{aligned} \tag{18.25}$$

$$\begin{aligned} \text{AR}(3): \text{Sales}_t &= 737.5405 + \underset{(.1616)}{.8987\ \text{sales}_{t-1}} - \underset{(.2220)}{.1082\ \text{sales}_{t-2}} \\ &\quad + \underset{(.1603)}{.1697\ \text{sales}_{t-3}} \\ R^2 &= .9698 \end{aligned} \tag{18.26}$$

[8] It should be noted that the exponential smoothing model of section 18.5 of the autoregressive models described herein are all special cases of *autoregressive integrated moving average* (*ARIMA*) models developed by Box and Jenkins. The Box-Jenkins approach, however, is beyond the scope of this text.

Table 18.10 Quarterly sales data for Johnson & Johnson (first quarter 2000 to fourth quarter 2010)

Quarter	S_t	S_{t-1}	S_{t-2}	S_{t-3}
2000Q1	7440			
2000Q2	7670	7440		
2000Q3	7438	7670	7440	
2000Q4	7298	7438	7670	7440
2001Q1	7855	7298	7438	7670
2001Q2	8179	7855	7298	7438
2001Q3	8058	8179	7855	7298
2001Q4	8225	8058	8179	7855
2002Q1	8743	8225	8058	8179
2002Q2	9073	8743	8225	8058
2002Q3	9079	9073	8743	8225
2002Q4	9403	9079	9073	8743
2003Q1	9821	9403	9079	9073
2003Q2	10333	9821	9403	9079
2003Q3	10454	10333	9821	9403
2003Q4	11254	10454	10333	9821
2004Q1	11559	11254	10454	10333
2004Q2	11484	11559	11254	10454
2004Q3	11553	11484	11559	11254
2004Q4	12752	11553	11484	11559
2005Q1	12832	12752	11553	11484
2005Q2	12762	12832	12752	11553
2005Q3	12230	12762	12832	12752
2005Q4	12610	12230	12762	12832
2006Q1	12992	12610	12230	12762
2006Q2	13363	12992	12610	12230
2006Q3	13157	13363	12992	12610
2006Q4	13682	13157	13363	12992
2007Q1	15037	13682	13157	13363
2007Q2	15131	15037	13682	13157
2007Q3	14910	15131	15037	13682
2007Q4	15957	14910	15131	15037
2008Q1	16194	15957	14910	15131
2008Q2	16450	16194	15957	14910
2008Q3	15921	16450	16194	15957
2008Q4	15182	15921	16450	16194
2009Q1	15026	15182	15921	16450
2009Q2	15239	15026	15182	15921
2009Q3	15081	15239	15026	15182
2009Q4	16551	15081	15239	15026
2010Q1	15631	16551	15081	15239
2010Q2	15330	15631	16551	15081
2010Q3	14982	15330	15631	16551
2010Q4	15644	14982	15330	15631

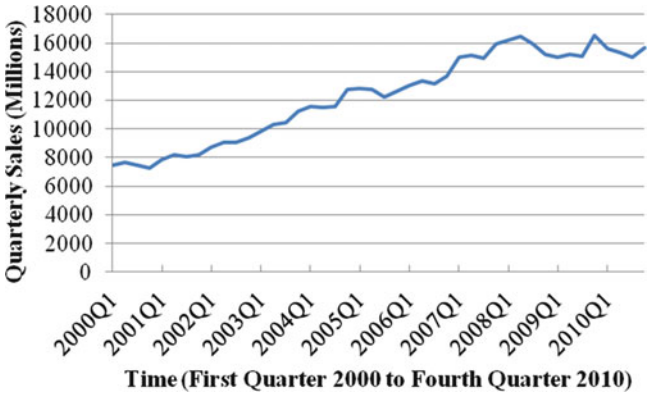


Fig. 18.16 Quarterly sales data for Johnson & Johnson

In Eqs. 18.24, 18.25, and 18.26, figures in parentheses under the coefficients are standard errors.

Table 18.10 makes it clear that the observations used to run AR(1), AR(2), and AR(3) are 43, 42, and 41, respectively. Therefore, by the central limit theorem, the parameter estimators divided by their standard errors approximate standard normal distributions.

From the standard error indicated in the parentheses and the parameter estimator, we can calculate the Z statistic for each regression slope. Looking up these Z statistics in Table A3 of Appendix A reveals that coefficients of sales_{t-1} in the AR(1), AR(2), and AR(3) model are significantly different from zero at the significance level of $\alpha = .05$. Hence, we conclude that the autoregressive processes can be used to forecast quarterly sales of Johnson & Johnson.

Substituting related quarterly sales data into the AR(1), AR(2), and AR(3) models, we obtain the following three alternative forecasted sales for the first quarter of 2011. Substituting $\text{sales}_{t-1} = 15{,}644$ into Eq. 18.24, we obtain the AR(1) forecast:

$$\begin{aligned}\text{Sales}_{2011Q1} &= 552.7913 + 0.9703(15644)\\ &= 15732\end{aligned}$$

Substituting $\text{sales}_{t-1} = 15{,}644$ and $\text{sales}_{t-2} = 14{,}982$ into Eq. 18.25, we obtain the AR(2) forecast:

$$\begin{aligned}\text{Sales}_{2011Q1} &= 586.6586 + .9106(15644) + .0580(14982)\\ &= 15700.72\end{aligned}$$

Substituting $\text{sales}_{t-1} = 15{,}644$, $\text{sales}_{t-2} = 14{,}982$, and $\text{sales}_{t-3} = 15{,}330$ into Eq. 18.26, we obtain the AR(3) forecast:

$$\begin{aligned}\text{Sales}_{2011Q1} &= 737.5405 + .8987(15644) - .1082(14982) + .1697(15330)\\ &= 15777.44\end{aligned}$$

To determine which model we should choose, we can use the mean absolute relative prediction error (MARPE) defined in Eq. 17.13 in Chap. 17 to see which one gives us the smallest error. Eq. 17.13 is

$$\text{MARPE} = \frac{|\hat{S}_t - S_t|}{S_t} \tag{17.13}$$

where $\hat{S}_t$ represents the sales forecast for time period t and S_t represents actual reported sales for time period t.

18.7 Summary

In this chapter, we examined time-series component analysis and several methods of forecasting. The major components of a time series are the trend, cyclical, seasonal, and irregular components. To analyze these time-series components, we used the moving-average method to obtain seasonally adjusted time series. After investigating the analysis of time-series components, we discussed several forecasting models in detail. These forecasting models are linear time trend regression, simple exponential smoothing, the Holt–Winters forecasting model without seasonality, the Holt–Winters forecasting model with seasonality, and autoregressive forecasting.

Many factors determine the power of any forecasting model. They include the time horizon of the forecast, the stability of variance of data, and the presence of a trend, seasonal, or cyclical component.

Questions and Problems

1. Consider a time series whose first value was recorded in December 1945. The last period for which there are records is June 1984.

   (a) How many full months of data are available?
   (b) How many full quarters of data are available?
   (c) How many full years of data are available?

2. Give an example of a time series you think may have

   (a) A moderately increasing linear trend
   (b) A decreasing linear trend
   (c) A curvilinear trend

3. The accompanying data indicate the number of mergers (x_t) that took place in a certain industry over a 15-year period.

Year	x_t	Year	x_t	Year	x_t
1970	15	1975	41	1980	148
1971	17	1976	85	1981	203
1972	24	1977	90	1982	249
1973	26	1978	110	1983	280
1974	30	1979	125	1984	307

(a) Plot these data on a frequency polygon.
(b) What type of trend (linear or nonlinear) might best be fitted to this time series?
(c) Is there evidence of seasonal variation in this series?

4. When a 5-month moving average is found for a time series, how many months do not have averages associated with them (a) at the beginning of the time series and (b) at the end of the time series?
5. Find the 3-year moving-average values for the merger time series described in question 3.
6. Find a 4-year moving-average series for the merger data given in question 3. Center the average on the years.
7. Use MINITAB to fit a least-squares trend line to the merger data given in question 3. Let $t = 1$ for 1970.
8. The following quarterly data show the number of cameras (in hundreds) returned to a particular manufacturer for warranty service over the past 5 years.

	Quarter			
Year	I	II	III	IV
5	.6	.4	.3	.6
4	.9	.6	.5	.8
3	1.6	1.8	1.8	1.6
2	1.3	1.1	1.0	1.3
1	1.5	1.3	1.1	1.5

Use MINITAB to answer the following questions:

(a) Plot this time series with time on the horizontal axis. Let $t = 1$ be the first quarter 5 years ago.
(b) Find the equation of the least-squares linear trend line that fits this time series. Let $t = 1$ be the first quarter 5 years ago.
(c) What would be the trend line for the second quarter of the current year—that is, 2 periods beyond the end of the actual date?

9. Determine the quarterly seasonal indexes for the warranty service time series described in question 8.
10. A cab company has supplied the accompanying data, which show the number of accidents involving its cabs over the past 5 years.

Year	Winter	Spring	Summer	Fall
5 years ago	7	5	4	6
4 years ago	7	7	5	7
3-years ago	11	10	6	9
2 years ago	22	11	7	10
Last year	16	12	9	12

Find the four seasonal indexes for accidents.

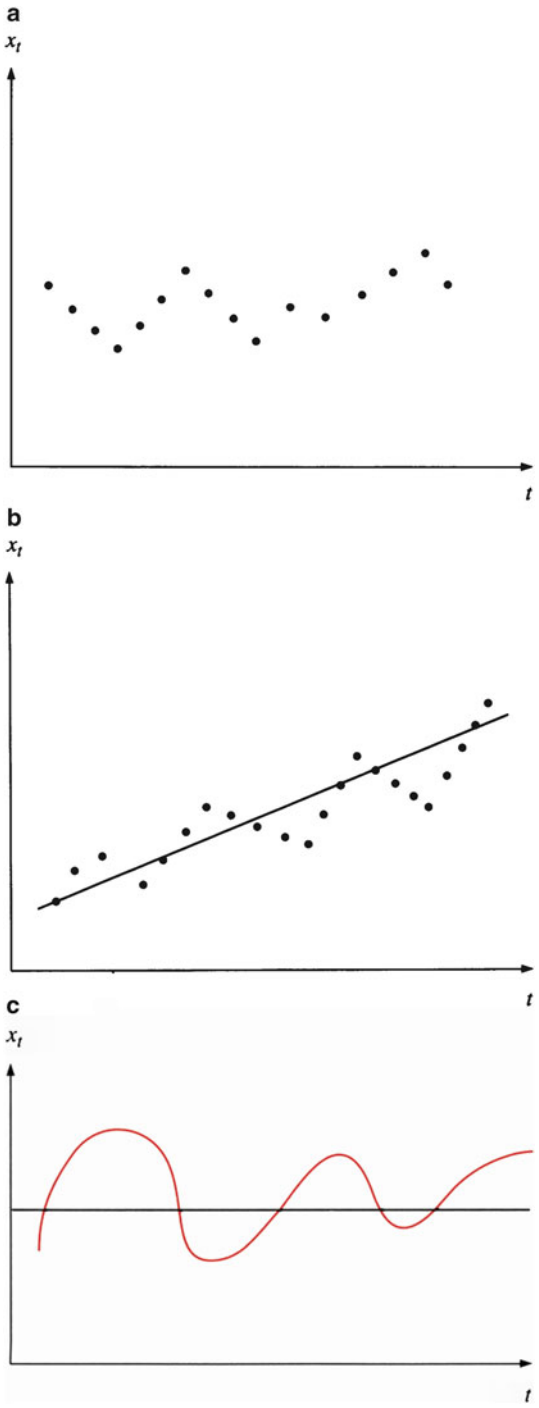
11. Actual billings for the Weygant Corporation were $135,478 in March, and the March seasonal index for this corporation's billings is 104. What is the seasonally adjusted March billing figure? What would be the expected annual billings based on the March figure?
12. The accompanying time series represents the number of patients received in a clinic emergency room. The seasonal indexes for each quarter are also given. Find the seasonally adjusted figures for the time series. Do these seasonal indexes tell the emergency room manager how many staff members to have on hand and what supplies to order for each quarter?

	Quarter			
	I	II	III	IV
Patient Visits	8,220	6,150	5,316	6,834
Seasonal index	115	73	85	110

13. What are time-series data? Why would we ever be interested in looking at time-series data? Give some examples of time-series data.
14. What is a seasonal factor? Why is seasonality sometimes a problem in modeling time-series data? Give some examples of seasonal effects.
15. Why do we sometimes need special techniques to analyze time-series data?
16. What is a business cycle? Why must businesses be able to forecast business cycles?
17. Define the four components of a time series.
18. Explain why it is easier to forecast when the time series contains seasonal effects rather than a cyclical effect?
19. Which of the components would you expect to exist in each of the following time series?

    (a) The quarterly earnings of Ford for the years 1981 through 1990
    (b) The monthly sales of Sears for 1990
    (c) The US unemployment rate for each year from 1981 through 1990
    (d) The US unemployment rate for each month in 1990

20. What are the advantages and disadvantages of using a simple moving-average technique for forecasting?
21. What are the advantages and disadvantages of using a linear trend for forecasting?
22. What are the advantages and disadvantages of using a nonlinear trend for forecasting?
23. What is exponential smoothing? What are the advantages and disadvantages of using exponential smoothing for forecasting?
24. What is an autoregressive process? What are the advantages and disadvantages of using an autoregressive process for forecasting?
25. What is the X-11 model? What is it used for? Briefly explain how the X-11 model is used in forecasting.
26. If you were asked to forecast the population of your town over the next 5 years, how would you do it? What information would you ask for?
27. Three time-series graphs follow. Try to identify the components of each time series.


a
x_t
t
b
x_t
t
c
x_t
t

28. Look at Fig. 18.1. What are the components of this time series? If you were asked to forecast this time series, what method would you use?
29. Look at Fig. 18.2. What are the components of this time series? If you were asked to forecast this time series, what method would you use?
30. Look at Fig. 18.3. What are the components of this time series? If you were asked to forecast the S&P 500 index, what method would you use?
31. If you were asked to forecast the size of the entering class at a college, how would you do this? What information would be useful in conducting your forecast?
32. You are told that the number of sales per month for a store follows an AR(1) process of the form

$$x_t = 125 + .6x_{t-1} + e_t$$

where e_t is normally distributed with zero mean and constant variance. Say $x_{30} =$ 1,000, $x_{31}=$ 1,125, and $x_{32} =$ 1,227. Forecast sales for the following time periods:

(a) x_{33}, x_{34}, and x_{35} at $t = 31$
(b) x_{34} and x_{35} at $t = 30$

33. You are told that the number of sales per month for a store follows an AR(2) process of the form

$$x_t = 15 + .6x_{t-1} - .2x_{t-2} + e_t$$

where e_t has a mean of zero and $E[e_t e_{t'}] = 0$ when $t \neq t'$. The values for the last three periods are $x_{101} = 823$, $x_{102} = 927$, and $x_{103} = 992$. Forecast sales for the next three periods $t =$ 104, 105, and 106.

Use the following information to answer questions 34–42. You are given the following return information for 3-month T-bills, the NYSE Index, Chrysler, Ford, and GM for the 3-year period from January 1985 through December 1987.

Month	T-Bill R_f	NYSE Index R_m	Chrysler R_t
85.01	.006280	.07950	.03906
85.02	.006687	.01661	.00376
85.03	.006886	−.00037	.05243
85.04	.006432	−.00277	−.00358
85.05	.006057	.05872	.02518
85.06	.005634	.01719	.03158
85.07	.005737	−.00351	−.00685
85.08	.005785	−.00463	.02414
85.09	.005753	−.03667	−.03030
85.10	.005801	.04462	.11189
85.11	.005865	.06884	.07233

(continued)

(continued)

Month	T-Bill R_f	NYSE Index R_m	Chrysler R_t
85.12	.005753	.04554	.09971
86.01	.005729	.00737	−.01072
86.02	.005721	.07375	.23035
86.03	.005321	.05560	.19273
86.04	.004920	−.01322	−.19220
86.05	.004993	.05147	.02759
86.06	.005041	.01509	.03020
86.07	.004736	−.05480	−.06229
86.08	.004495	.07312	.08392
86.09	.004237	−.07957	−.06193
86.10	.004213	.05402	.06944
86.11	.004350	.01857	.02273
86.12	.004495	−.02677	−.05143
87.01	.004414	.12823	.29054
87.02	.004543	.04100	−.01309
87.03	.004543	.02469	.18037
87.04	.004583	−.01483	.03846
87.05	.004599	.00644	−.11111
87.06	.004607	.04797	.01103
87.07	.004623	.04682	.19414
87.08	.004904	.03688	.09816
87.09	.005193	−.02085	−.06983
87.10	.004977	−.21643	−.35649
87.11	.004623	−.07547	−.23944
87.12	.004688	.06851	.10494

Month	Ford R_2	GM R_3
85.01	.07945	.06061
85.02	−.08461	−.02857
85.03	−.05042	−.08176
85.04	−.02124	−.07363
85.05	.06422	.07763
85.06	.03736	.00524
85.07	.00222	−.01736
85.08	−.01401	−.03003
85.09	.00568	−.00557
85.10	.06667	−.00373
85.11	.16129	.06929
85.12	.07407	.03084
86.01	.09181	.05151
86.02	.14571	.06757
86.03	.14111	.10932
86.04	−.06626	−.07246
86.05	.06446	.01250
86.06	.02717	−.02665

(continued)

(continued)

Month	Ford R_2	GM R_3
86.07	−.01950	−.12238
86.08	.11682	.07523
86.09	−.11297	−.05903
86.10	.09481	.04981
86.11	.01961	.04218
86.12	−.03846	−.09434
87.01	.33378	.14015
87.02	.02689	.00831
87.03	.10475	.04690
87.04	.08741	.15200
87.05	−.00137	−.03889
87.06	.08941	−.03079
87.07	.03409	.07564
87.08	.06273	.04923
87.09	−.09259	−.09783
87.10	−.21939	−.29518
87.11	−.05795	−.01496
87.12	.05975	.08869

34. Use MINITAB to plot the return data for T-bills against time. (Let $t = 1$ be the first month.) Can you identify any of the components of the time series?
35. Compute a simple 3-period moving average for the return on T-bills. Forecast the value for January 1988 using this method.
36. With the MINITAB program, use an AR(1) model to describe the time-series behavior of T-bills. Forecast the value for January 1988 using the AR(1) procedure.
37. Using only data from January 1985 through November 1987, forecast the value for December 1987, using both the 3-period moving average and the AR(1) model. Compare your results. Which model forecasts better?
38. Repeat question 37 using the data for the NYSE index.
39. Repeat question 37 using the data for Chrysler.
40. Repeat question 37 using the data for Ford.
41. Repeat question 37 using the data for GM.
42. Compare the two methods you used for forecasting in questions 34–41. Is one method superior to the other in all cases?
43. Suppose you are an investment analyst and are interested in estimating the future dividend for Hamby Corp. You know that Hamby's dividends grow at an exponential rate—that is,

$$D_t = D_0(1 + g)^t$$

where D_t is the dividend in year t, D_0 is the dividend this year, and g is the growth rate of dividends (assumed to be constant). Is there any way to transform this model into a linear regression?

44. Suppose you are given the following dividend information for Hamby Corp. Forecast the dividend for years 6, 7, 8, 9, and 10, using the method you proposed in question 43.

Year	*D*
0	1.25
1	1.32
2	1.37
3	1.45
4	1.53
5	1.60

45. Again use the data given in question 44, but this time apply a linear time trend. Plot the estimates from this regression and from your results in question 44.
46. Suppose you have the following information about a company's EPS. What would be the best method for modeling this company's EPS? Forecast the EPS for years 6, 7, 8, 9, and 10.

Year	EPS
0	$3.25
1	3.65
2	4.03
3	4.45
4	4.87
5	5.09

47. Explain why we use t as an explanatory variable in a linear time trend model when it is not time that causes the dependent variable to change.
48. Suppose you are given the following sales information for Julian Corp. Estimate the growth rate of sales for Julian Crop. Use this information to forecast the company's sales for year 10.

Year	Sales
0	1,250,625
1	1,321,001
2	1,372,435
3	1,458,020
4	1,531,035
5	1,600,995

49. Evaluate the following statement: "Because sales have increased at a steady rate over the last 10 years, the best way to forecast future sales is to use a linear time trend."

50. Go to the library and obtain the earnings per share for General Motors for the years 1979 through 1988. Use the data for earnings in 1979 through 1988 to obtain a forecasting equation.
51. Indicate which component of a time series will be affected by each of the following events:

    (a) A hurricane that results in the postponement of consumer purchases
    (b) A downturn in business activity
    (c) The annual Columbus Day sale at a department store
    (d) A flood at a wholesale warehouse that results in a delay in the shipment of clothing to a local department store
    (e) A general increase in the demand for video cameras

52. You are given the following sales information (in millions of dollars) on Acme Widget Company:

Year	Sales ($)	Year	Sales ($)
1985	3.2	1989	4.8
1986	4.5	1990	5.1
1987	3.9	1991	5.6
1988	4.2		

    (a) Use a line chart to graph sales.
    (b) Estimate the relationship between sales and time, using a time trend regression.

    Use the following information on total nonfarm payrolls in New Jersey from 1965 to 1989, which is taken from *New Jersey Economic Indicators,* March 1990, to answer questions 53–57.

Year	Total nonfarm Payrolls	Year	Total nonfarm Payrolls
1965	2,257.8	1978	2,961.9
1966	2,359.1	1979	3,027.2
1967	2,421.5	1980	3,060.4
1968	2,485.2	1980	3,060.4
1969	2,569.6	1981	3,089.9
1970	2,606.2	1982	3,092.7
1971	2,607.6	1983	3,165.1
1972	2,674.4	1984	3,329.3
1973	2,760.8	1985	3,414.1
1974	2,783.4	1986	3,489.9
1975	2,699.9	1987	3,581.6
1976	2,753.7	1988	3,659.5
1977	2,836.9	1989	3,709.8

53. Use the MINITAB program to plot the data for nonfarm income, and identify the components of the time series.
54. Compute the 3-year moving average for nonfarm income. Use this information to forecast nonfarm income in 1990 and in 1991.
55. Use the MINITAB program to do a time trend regression to forecast nonfarm income in 1990 and in 1991.
56. Use a first-order autoregressive process to forecast nonfarm income in 1990 and in 1991.
57. Compare the different forecasts of nonfarm income that you made in questions 54–56.

Use the following employment data (in thousands) for the United States and for New Jersey to answer questions 58–65.

	Employment	
Year	United States	New Jersey
1970	78,678	2,859
1971	79,367	2,840
1972	82,153	2,935
1973	85,064	3,011
1974	86,794	3,023
1975	85,846	2,929
1976	88,752	2,973
1977	92,017	3,065
1978	96,048	3,209
1979	98,824	3,323
1980	99,303	3,334
1981	100,397	3,330
1982	99,526	3,306
1983	100,834	3,385
1984	105,005	3,589
1985	107,150	3,621
1986	109,597	3,712
1987	112,440	3,806
1988	114,968	3,824
1989	117,342	3,826

58. Graph the employment for the United States, and try to identify the components of the time series.
59. Compute the 4-year moving average for employment in the United States. Use this information to forecast employment in the United States in 1990.
60. Use a time trend regression to forecast employment in the United States in 1990, 1991, and 1992.
61. Use a first-order autoregressive model to forecast employment in the United States in 1990, 1991, and 1992.
62. Do you think the first-order AR(1) is a good model to use to explain the data?

63. Compare the different forecasts generated for 1990 by the methods you used in questions 59–62. Which method do you think is best? Why?
64. Plot the New Jersey employment data. Do you think the linear trend model provides a good approximation of the data? Use the data to forecast the employment in 1990.
65. Compare your forecasts for New Jersey with your forecasts for the United States. Which set of data is harder to forecast? Why?
    Use the following data on the labor force in thousands of people in the United States and in New Jersey to answer questions 66–70.

	Labor Force	
Year	United States	New Jersey
1970	82,771	2,996
1971	84,382	3,012
1972	87,034	3,117
1973	89,429	3,190
1974	91,949	3,226
1975	93,775	3,264
1976	96,158	3,318
1977	99,009	3,383
1978	102,251	3,457
1979	104,962	3,570
1980	106,940	3,594
1981	108,670	3,593
1982	110,204	3,632
1983	111,550	3,673
1984	113,544	3,825
1985	115,461	3,839
1986	117,834	3,908
1987	119,865	3,966
1988	121,669	3,975
1989	123,869	3,989

66. Plot the labor force in the United States and in New Jersey, and try to identify the components of the time series. Which labor force data appear to be more stable?
67. Compute the 5-year moving averages for the labor force in the United States and in New Jersey.
68. Use a linear time trend regression to estimate the labor force in the United States and in New Jersey in 1990, 1991, and 1992.
69. Use an exponential trend model to forecast the labor force in the United States and in New Jersey for 1990–1993.
70. What are the growth rates of the United States and New Jersey labor forces? Does the linear model or the exponential trend model give a faster growth estimate?
71. Suppose you generate the following data by tossing a coin 50 times. Let the initial value be \$50. If you toss a head, increase the value by \$.50. If you toss a

tail, decrease the value by $.50. Graph the data. Does this series of data exhibit any time-series pattern? What time-series pattern would you expect it to exhibit?

72. Can you use any regression or time-series method to forecast the values in periods 50, 51, and 52 in question 71?
73. What is the best forecast for the value at period 51?
74. Suppose you adjusted the data generated in question 71 by adding $.25 to every fourth coin toss. Graph these data. Does this new series exhibit any time-series pattern? What time-series pattern would you expect it to exhibit?
75. What is the best forecast for the time series generated in question 74?
76. Johnson & Johnson's quarterly sales, in millions of dollars, from first quarter 1990 to first quarter 1991 are

First quarter 1990	2,809
Second quarter 1990	2,825
Third quarter 1990	2,775
Fourth quarter 1990	2,794
First quarter 1991	3,149

Use this set of data and the data in Table 18.10 to run an autoregression model with 1, 2, and 3 lags from first quarter 1980 to fourth quarter 1990. Use MINITAB. Then use actual sales data for first quarter 1991 to calculate the prediction error as defined in Eq. 17.13.

77. The contents in the last column of Table 18.5 are the adjusted EPS of Johnson & Johnson by the Seasonal Index from first quarter 2000 to fourth quarter 2010. Denote them as x_t^*.

    (a) Determine the linear trend between x_t^* and t.
    (b) Determine the quadratic trend between x_t^* and t.
    (c) Determine the cubic trend between x_t^* and t.

78. (Problem *77* continued.) Forecast the adjusted or deseasonalized EPS of the Johnson & Johnson from the first quarter 2011 to fourth quarter 2011 by using the linear trend estimates, quadratic trend estimates, and cubic trend estimates, respectively.
79. (Problem *78* continued.) Find the forecasts of EPS of the Johnson & Johnson from the first quarter 2011 to fourth quarter 2011.
80. Use a fourth-order autoregressive model to forecast EPS of the Johnson & Johnson for the first quarter of 2011.
81. Use the exponential smoothing method, with the smoothing constant α being 0.1, and 0.9, respectively, to forecast the return of NYSE Index. Which smoothing constant has smaller MSE?

Appendix 1: The Holt–Winters Forecasting Model for Seasonal Series

In this appendix, we will generalize the Holt–Winters forecasting model discussed in Sect. 18.5 to take into account the existence of seasonality. As in the nonseasonal case, we will use x_t, s_t, and T_t to denote, respectively, the observed value and the level and trend estimates at time t. F_t is used to denote the seasonal factor, so if the time series contains L periods per year, the seasonal factor for the corresponding period in the previous period will be F_{t-L}. The Holt–Winters method for seasonal series can be expressed by the following three equations:

$$s_t = \alpha\left(\frac{x_t}{F_{t-L}}\right) + (1-\alpha)(s_{t-1} + T_{t-1}) \tag{18.27}$$

$$T_t = \beta(s_t - s_{t-1}) + (1-\beta)T_{t-1} \tag{18.28}$$

$$F_t = \gamma\left(\frac{x_t}{s_t}\right) + (1-\gamma)F_{t-L}. \tag{18.29}$$

where α, β, and γ are smoothing constants whose values are set between 0 and 1.

In Eq. 18.27, the term $s_{t-1} + T_{t-1}$ represents an estimate of the level at time t, formed 1 time period earlier. This estimate is updated when the new observation x_t becomes available. However, here it is necessary to remove the influence of seasonality from that observation by deflating it by the latest available estimate, F_{t-L}, of the seasonal factor for that period. The updating equation for trend, Eq. 18.28, is identical to that used previously, Eq. 18.19b, in the text.

Finally, the seasonal factor is estimated by Eq. 18.29. The most recent estimate of the factor, available from the previous year, is F_{t-L}. However, dividing the new observation x_t by the level estimate s_t suggests a seasonal factor x_t/s_t. The new estimate of the seasonal factor is then a weighted average of these two quantities.

The procedure for forecasting via the Holt–Winters forecasting model for seasonal series is similar to that for nonseasonal series. Here, the forecast for a particular month includes the effect of all three smoothing equations. The forecast for m periods ahead is

$$\hat{x}_{t+m} = (s_t + mT_t)(F_{t+m-L}) \tag{18.30}$$

If no seasonality exists—that is, if $F_{t+m-L} = 1$—then this equation reduces to Eq. 18.20 in the text.

We will use quarterly data listed in Table 18.4 in the text for Johnson & Johnson (J&J) during the period first quarter 2000 through fourth quarter 2010 to demonstrate how Eqs. 18.27, 18.28, 18.29, and 18.30 are used to do exponential smoothing and forecasting.

Table 18.11 Seasonal Index and seasonally adjusted EPS for J&J in terms of the first 12 quarters' data

(1) Date	(2) x_t	(3) z_t^*	(4) x_t/z^*	(5) Seasonal index	(6) Seasonally adjusted EPS, d_t
2000.1	0.86			0.503173	1.709154
2	1.8			0.750721	2.397696
3	2.68	2.1775	1.230769	1.207303	2.219824
4	3.3	2.095	1.575179	1.538804	2.144523
2001.1	1	1.84875	0.540906	0.503173	1.987389
2	1	1.52375	0.656276	0.750721	1.332053
3	1.51	1.295	1.166023	1.207303	1.250722
4	1.87	1.26375	1.479723	1.538804	1.21523
2002.1	0.6	1.31	0.458015	0.503173	1.192433
2	1.15	1.37875	0.834089	0.750721	1.531861
3	1.73			1.207303	1.432946
4	2.2			1.538804	1.429682

Table 18.12 Calculation of seasonal indexes of EPS for J&J

	Quarter				
Year	*1*	*2*	*3*	*4*	Sums
2000			1.231	1.575	
2001	0.541	0.656	1.166	1.480	
2002	0.458	0.834			
Median	0.499	0.745	1.198	1.527	3.970
Seasonal Index	0.503	0.751	1.207	1.539	4.000

Example 18.7 The Holt–Winters Forecasting Model for J&J's Quarterly EPS. Table 18.4 and Fig. 18.6 in the text make it clear that Johnson & Johnson's quarterly EPS in the period 2000–2010 exhibited significant seasonality. The fourth-quarter EPS especially appeared to be considerably higher than those for the other three quarters.

The Holt–Winters forecasting model with seasonality is used to determine the smoothed value, s_t, and the predicted value, $\hat{x}$, for each time period. The smoothing constants are $\alpha = .2$, $\beta = .3$, and $\gamma = .3$.

First, we use the first three years of data to determine the seasonal indexes. Working with Eq. 18.9, we present the percentage of moving average (PMA) in terms of the first three years' data in column (4) of Table 18.11. Table 18.12 shows the procedure for calculating the seasonal index in terms of the first three years' data. These indexes are

$$\text{Quarter } 1 = 0.503 \qquad \text{Quarter } 2 = 0.751$$
$$\text{Quarter } 3 = 1.207 \qquad \text{Quarter } 4 = 1.539$$

and these are the four values of F_t in 1999.

The data from the first three years were seasonally adjusted to obtain d_t; see column (6) of Table 18.11. Drawing a least-squares line through these 12 values by means of simple time trend linear regression produces

$$\hat{d}_t = 2.192966 - .08298t$$

The value $\hat{b} = .08298$ becomes the initial trend estimate of T_0. Finally, the initial smoothed value for fourth quarter 1999 is

$$\begin{aligned} s_0 &= [a + b(0)](\text{initial seasonal index for fourth quarter}) \\ &= (2.192996)(1.539) = 3.374543 \end{aligned}$$

This estimate of s_0 becomes the forecast value for each of the quarters in 2000, as indicated in column (6) of Table 18.13.

The calculation of Table 18.13 in terms of $t = 10$ is shown as follows:

1. $x_{10} = 1.15$
2. Substituting related information into Eq. 18.27 yields

$$\begin{aligned} s_{10} &= .2\left(\frac{x_{10}}{F_{10-4}}\right) + .8(s_9 + T_9) \\ &= .2\left(\frac{x_{10}}{F_6}\right) + .8(s_9 + T_9) \\ &= .2\left(\frac{1.15}{0.678411}\right) + .8(1.042416 - .22779) \\ &= .99073 \end{aligned}$$

3. Substituting related information into Eq. 18.28 yields

$$\begin{aligned} T_{10} &= .3(s_{10} - s_9) + .7T_9 \\ &= .3(.99073 - 1.042416) + .7(-0.22779) \\ &= -0.17496 \end{aligned}$$

4. Substituting related information into Eq. 18.29 yields

$$\begin{aligned} F_{10} &= .3\left(\frac{x_{10}}{s_{10}}\right) + .7F_6 \\ &= .3\left(\frac{1.15}{.99073}\right) + .7(.678411) \\ &= .823116 \end{aligned}$$

Similarly, we can calculate all other values of s_t, T_t, and F_t, which are listed in columns (5), (3), and (4), respectively. Figure 18.17 presents actual data and smoothed data s_t.

Using Eq. 18.30, we estimate $\hat{x}_{t+1}$ ($t = 5,6,\ldots,44$); it is shown in column (6) of Table 18.13. For example,

Table 18.13 Solution using Holt–Winters model with seasonality ($\alpha = .2$, $\beta = .3$, $\gamma = .3$)

t	x_t	T_t	F_t	s_t	$\hat{x}_t$	$x_t - \hat{x}_t$
			0.503173			
			0.750721			
			1.207303			
		−0.08298	1.538804	3.374543		
1	0.86	−0.17792	0.438941	2.975085	1.656227	−0.79623
2	1.8	−0.20189	0.724233	2.717271	2.09989	−0.29989
3	2.68	−0.21962	1.172438	2.456271	3.03683	−0.35683
4	3.3	−0.22515	1.523465	2.218224	3.441764	−0.14176
5	1	−0.20804	0.453593	2.050102	0.874843	0.125157
6	1	−0.23572	0.678411	1.749802	1.334081	−0.33408
7	1.51	−0.24929	1.129111	1.46885	1.775169	−0.26517
8	1.87	−0.24881	1.525832	1.221142	1.857959	0.012041
9	0.6	−0.22779	0.490191	1.042416	0.441041	0.158959
10	1.15	−0.17496	0.823116	0.99073	0.552653	0.597347
11	1.73	−0.13197	1.331536	0.959054	0.921098	0.808902
12	2.2	−0.09509	1.762796	0.950032	1.261986	0.938014
13	0.7	−0.0607	0.559727	0.969558	0.419086	0.280914
14	1.09	−0.03578	0.905841	0.991931	0.748093	0.341907
15	1.8	−0.01204	1.453671	1.035285	1.273149	0.526851
16	2.42	0.008934	1.898087	1.09316	1.803772	0.616228
17	0.84	0.032852	0.605039	1.181821	0.616872	0.223128
18	1.67	0.070587	1.007842	1.340457	1.100301	0.569699
19	2.46	0.087461	1.520538	1.467289	2.051194	0.408806
20	2.87	0.084899	1.885506	1.54621	2.951051	−0.08105
21	0.96	0.082233	0.601062	1.622222	0.986885	−0.02688
22	1.83	0.088911	1.023434	1.726716	1.717821	0.112179
23	2.68	0.085726	1.509804	1.805008	2.760729	−0.08073
24	3.38	0.079839	1.861778	1.871112	3.564991	−0.18499
25	1.11	0.073586	0.593272	1.930107	1.172642	−0.06264
26	2.07	0.07472	1.025748	2.007475	2.050647	0.019353
27	3.01	0.069407	1.49426	2.064483	3.143705	−0.13371
28	3.76	0.062548	1.837582	2.111027	3.97283	−0.21283
29	0.89	0.022143	0.546244	2.038891	1.289522	−0.39952
30	1.95	0.012544	1.006337	2.029037	2.1141	−0.1641
31	2.84	0.004085	1.46915	2.013386	3.050653	−0.21065
32	3.67	0.002868	1.833139	2.013415	3.707269	−0.03727
33	1.27	0.021389	0.565719	2.07802	1.101383	0.168617
34	2.45	0.041499	1.043702	2.166442	2.112714	0.337286
35	3.64	0.05768	1.51119	2.261878	3.243796	0.396204
36	4.62	0.069723	1.870561	2.359699	4.252072	0.367928
37	1.27	0.058653	0.555249	2.392524	1.374369	−0.10437
38	2.43	0.051278	1.031013	2.426592	2.558299	−0.1283
39	3.64	0.047127	1.501008	2.464035	3.744531	−0.10453
40	4.45	0.039196	1.846676	2.484723	4.697282	−0.24728
41	1.64	0.064978	0.57719	2.609861	1.401404	0.238596
42	2.89	0.072672	1.042762	2.700485	2.757794	0.132206
43	4.14	0.071771	1.499056	2.770155	4.162532	−0.02253
44	4.85	0.058836	1.812537	2.798809	5.248116	−0.39812

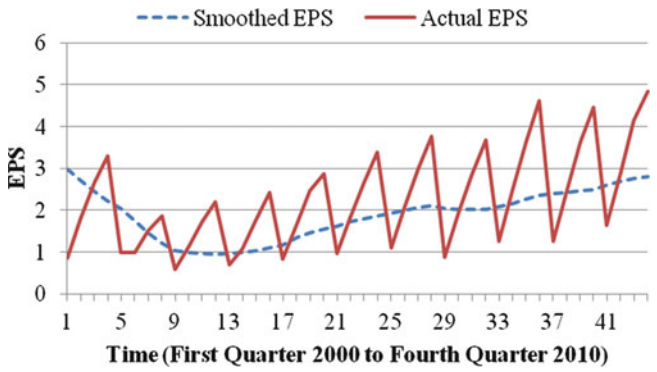


Fig. 18.17 Quarterly earnings per share of J&J (actual and smoothed EPS)

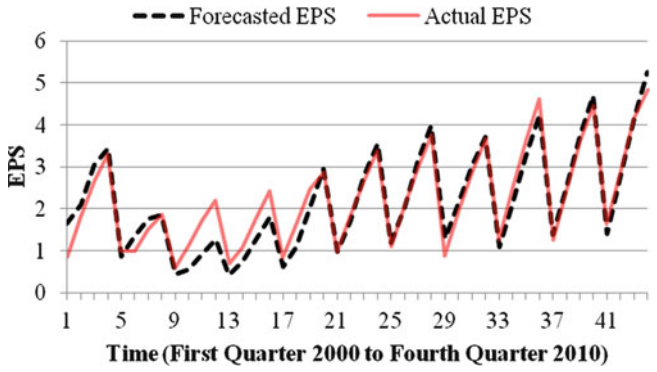


Fig. 18.18 Quarterly earnings per share of J&J (actual and forecasted EPS)

$$\hat{x}_{11} = (s_{10} + T_{10})(F_7) = (.99073 - .17496)(.678411)$$
$$= .921098$$

Figure 18.18 presents actual data and forecasted data ($\hat{x}_t$). If we let $m \geq 1$, then we can forecast future observations. For example, to forecast the EPS of J&J in the third quarter of 2002, we let $m = 1$. Finally, in the last column of Table 18.13, we present the residual in period t, $(x_t - \hat{x}_t)$.

Chapter 19
Index Numbers and Stock Market Indexes

Chapter Outline

Key Terms

Index number	Quantity index
Simple aggregative price index	Paasche quantity index
Price relative	Fisher's ideal quantity index
Consumer price index	FRB index of industrial production
Base year	Value index
Simple relative price index	Stock market index
Weighted relative price index	Market-value-weighted index
Laspeyres price index	S&P 500 index
Weighted aggregative price index	Price-weighted index
Paasche price index	Dow Jones Industrial Average
GNP deflator	Equally weighted index
Fisher's ideal price index	Indicators of economic activity
Option on stock index	Option on currency
Index futures	Hedge ratio

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5_19,

19.1 Introduction

Business executives and government officials often make judgments that involve summarizing how business, economic, and financial variables change with time or place. Examples of variation over *time* include variation in gross national production, variation in the price of consumer goods, and variation in stock market prices, As an example of variation with changes in *place*, consider a company that wishes to transfer an executive from Chicago to San Francisco. What should be the executive's minimum salary increase to compensate for the higher cost of living in San Francisco?

In all these cases, we need to have a single composite figure to summarize the average difference between two time periods or between the two cities. Index numbers can be used to answer questions of this type. An *index number* is a summary measure that compares related items over time or place. In other words, index numbers enable us to express the level of an activity or phenomenon in relation to its level at another time or place.

In this chapter we will investigate how alternative index numbers are compiled and used in business, economics, and finance analyses. First, a discussion of price indexes, quantity indexes, and value indexes lays the foundation for an understanding of economic and financial indexes. Then we develop several types of stock market indexes and examine the major indexes provided in the daily financial news. Finally, applications of index numbers in business and economics are discussed.

19.2 Price Indexes

In this section we first develop a *simple aggregative price index* based on a single good and then expand the concept to a combination of several goods. We also address some of the problems associated with price indexes and explore techniques that have been developed to deal with these problems.

19.2.1 Simple Aggregative Price Index

In its simplest form, an index number is nothing more than a percentage figure that expresses the relationship between two numbers, one of the numbers being used as the base. For example, in a time series of prices of a particular commodity, we can express the prices as percentages by dividing each figure by the price in the base period. These percentages are referred to as *price relatives*.

An understanding of the simple aggregative price index can perhaps best be facilitated through the use of an example of price relatives. Assume that the price of eggs has risen over three consecutive years as follows:

1989: $P_0 = \$1.00$
1990: $P_1 = \$1.20$
1991: $P_2 = \$1.50$

To illustrate the change in prices, we calculate the ratio of prices with respect to a *base year*, the year from which the future price changes are measured. The appropriate base year depends on relevant economic factors. The base years of US government indexes are shifted forward approximately every decade to reflect changes in economic conditions over time. In our example, the base year is 1989.

1989: $P_0/P_0 = 1.00/1.00 = 1.00 = I_0$
1990: $P_1/P_0 = 1.20/1.00 = 1.20 = I_1$
1991: $P_2/P_0 = 1.50/1.00 = 1.50 = I_2$

$I_1 = 1.20$ means that the price of eggs increased 20 % between 1989 and 1990. Likewise, $I_2 = 1.50$ means that between 1989 and 1991 there was a 50 % increase in the price of eggs. Because 1989 is the base year, I_0 must equal 1.00.

Government price indexes are usually expressed on a basis of 100, and to be consistent with this practice, we will multiply each of the foregoing indexes by 100. Hence, the price indexes for eggs from 1989 to 1991 are

I_0: $1.00 \times 100 = 100$
I_1: $1.20 \times 100 = 120$
I_2: $1.50 \times 100 = 150$

Each price of eggs is a price relative – the ratio of the price in a given year to the price in the base year.

The previous example of a price index was expressed in terms of only one commodity, eggs. A more realistic approach would be to include a group of commodities, as does the *consumer price index* (CPI). The CPI measures the cost of a market basket of some 2,000 consumer goods and services purchased by a "typical" urban family. The composition of this basket is food, clothing, housing, fuels, transportation, and medical care. For the sake of simplicity and ease of understanding, our market basket will consist of just four commodities:

One dozen eggs	One pound of butter
One gallon of milk	One loaf of bread

We will also assume that the same amount of each good was purchased in each of our three consecutive years. Table 19.1 illustrates the prices, for the years 1989–1991, of the individual goods that make up our market basket.

The table indicates that the same market basket of commodities cost $4.00 in 1989, $5.00 in 1990, and $6.00 in 1991. However, because we are interested in determining the price index of the market basket for the various years, we simply calculate the price relatives (ratios of prices between two different periods) of the group of commodities, using 1989 as the base year. The price indexes for 1989–1991 are

Table 19.1 Price and quantity for four commodities, 1989–1991

Commodity	Quantity	1989 P_0	1990 P_1	1991 P_2
Eggs	One dozen	1.00	1.20	1.50
Milk	One gallon	1.50	1.75	2.00
Butter	One pound	1.10	1.35	1.60
Bread	One loaf	0.40	0.70	0.90
		4.00	5.00	6.00

1989: $4.00/4.00 \times 100 = 100$
1990: $5.00/4.00 \times 100 = 125$
1991: $6.00/4.00 \times 100 = 150$

The indexes indicate that the cost of this *specific* list of goods increased 25 % between 1989 and 1990 and 50 % between 1989 and 1991.

Formally, we can write the simple price index as follows:

$$I_t = \frac{\sum_{i=1}^{4} P_{ti}}{\sum_{i=1}^{4} P_{0i}} \times 100 \tag{19.1}$$

where

I_t = price index for the year t
P_{ti} = price of the ith commodity in the year t
P_{0i} = price of the ith commodity in the base year 0

Note that the quantity for the ith good can be regarded as $Q_{0i} = 1$ for all i.

19.2.2 Simple Average of Price Relatives

Two disadvantages of the simple aggregate price index give rise to a need for the simple average of relatives. These two disadvantages are:

1. The units used to state the prices of the commodities affect the price index. For example, if the price of eggs were stated in half dozens rather than in dozens, then the price indexes would be

1989: $3.50/3.50 \times 100 = 100$
1990: $4.40/3.50 \times 100 = 125.7$
1991: $5.25/3.50 \times 100 = 150$

These indexes are not identical to those calculated from the price of eggs per dozen.

Table 19.2 Calculation of the simple average of price relatives

Commodity	1989 P_0/P_0	1990 P_1/P_0	1991 P_2/P_0
Eggs	100	$1.20/1.00 \times 100 = 120$	$1.50/1.00 \times 100 = 150$
Milk	100	$1.75/1.5 \times 100 = 117$	$2.00/1.50 \times 100 = 133$
Butter	100	$1.35/1.10 \times 100 = 123$	$1.60/1.10 \times 100 = 145$
Bread	100	$0.70/0.40 \times 100 = 175$	$0.90/0.40 \times 100 = 225$
Shirt	100	$20.00/16.00 \times 100 = 125$	$10.00/16.00 \times 100 = 63$
	500 %	660 %	716 %

2. The index does not consider the relative importance of the commodities, and it is unduly influenced by the price variation of high-priced commodities. For example, if our market basket were enlarged to include a shirt that cost \$16 in 1989, \$20 in 1990, and \$10 in 1991, calculating our price indexes in terms of Eq. 19.1 would yield

$$I_{89} = (4.00 + 16.00)/(4.00 + 16.00) \times 100 = 20/20 \times 100 = 100$$
$$I_{90} = (5.00 + 20.00)/(4.00 + 16.00) \times 100 = 25/20 \times 100 = 125$$
$$I_{91} = (6.00 + 10.00)/(4.00 + 16.00) \times 100 = 16/20 \times 100 = 80$$

The shirt makes up a majority of the index. The decrease in the price of a shirt makes the index decrease by 20 %, even though that shirt is not the most important item for consumers.

Because of these limitations, we need an index that removes the bias due to the difference in measurement and takes the relative importance of the commodity into account. We can improve on our index by taking an average of the price relatives.

Using the data from Table 19.2, we can calculate the *simple relative price index* in period t as follows:

$$I_t = \frac{\sum_{i=1}^{n} (P_{ti}/P_{0i} \times 100)}{n} \tag{19.2}$$

Hence, the averages of the price relatives for 1990 and 1991 are 660/5, or 132, and 716/5, or 143.2, respectively.

Each item in the index is weighted by $1/P_0$, which makes all items equally important. Thus, we have removed the influence of different units of measurement for the various commodities. However, this kind of index still doesn't take the relative importance of the commodity into account.

19.2.3 Weighted Relative Price Index

One major disadvantage of the simple relative price index is that it treats all commodities as equal. An index should reflect the value of some commodities in

relation to the value of others. We need a *weighted relative price index* – one that takes into consideration the weights, or worth, of various commodities. We base the value of commodity weights on the quantity purchased. In other words, the total dollars spent on the commodities determine their weight in the index.

Suppose the following are the amounts spent on our market basket, V_{0i}, and the related prices and quantities of those commodities, in 1989.

Commodity	V_{0i}	P_{0i}	Q_{0i}
Eggs	\$150	\$1.00	150
Milk	450	1.50	300
Butter	220	1.10	200
Bread	440	0.40	1,100
Shirts	160	16.00	10
	\$1,420		

That is, the value of the ith commodity purchased in 1989 is

$$V_{0i} = P_{0i} \times Q_{0i} \tag{19.3}$$

Hence, Q_{0i} is the quantity of eggs purchased in the base year (1989).

Accordingly, the purpose of the weighted relative price index is to show how much we need to spend in subsequent years to buy the same amount of commodities as we bought in the base year.

Formally, we derive the weighted relative for the ith commodity in period t as follows:

$$\left(\frac{P_{ti}}{P_{0i}}\right)V_{0i} = \left(\frac{P_{ti}}{P_{0i}}\right)P_{0i} \times Q_{0i} = P_{ti} \times Q_{0i} \tag{19.4}$$

Summing this for the five commodities in our market basket gives us the price index in period t, I_t, as

$$I_t = \frac{\sum_{i=1}^{5}(P_{ti}/P_{0i})V_{0i}}{\sum_{i=1}^{5}V_{0i}} \times 100 \tag{19.5}$$

Table 19.3 shows the individual weighted price relatives for 1989–1991. Thus, the weighted relative price indexes are

1989: (1,420/1,420)(100) = 100
1990: (1,948/1,420)(100) = 137
1991: (2,234/1,420)(100) = 157

Table 19.3 Calculation of individual weighted price relatives for 1989–1991

Commodity	Weight	$(P_{0i}/P_{0i})\ V_{0i}$	$(P_{1i}/P_{0i})\ V_{0i}$	$(P_{2i}/P_{0i})\ V_{0i}$
Eggs	150	150	(1.2)150 = 180	(1.5)350 = 225
Milk	450	450	(1.17)450 = 527	(1.33)450 = 599
Butter	220	220	(1.23)220 = 271	(1.45)220 = 319
Bread	440	440	(1.75)440 = 770	(2.25)440 = 990
Shirts	*160*	*160*	(1.25)160 = *200*	(0.63)160 = *101*
	1,420	1,420	1,948	2,234

This weighted relative price index is higher than the simple relative price index because bread has a high importance rating (that is now taken into account) and because bread underwent a substantial price increase.

19.2.4 Weighted Aggregative Price Index

In this section we discuss three weighted aggregative price indexes: the Laspeyres price index, the Paasche price index, and Fisher's ideal price index.

19.2.4.1 The Laspeyres Price Index

We can rewrite the weighted price index given in Eq. 19.5 by using Eq. 19.4.

$$\begin{aligned} I_t &= \frac{\sum_{i=1}^{5}\left(\frac{P_{ti}}{P_{0i}}\right)V_{0i}}{\sum_{i=1}^{5}V_{0i}} \times 100 \\ &= \frac{(P_{t1}/P_{01})(P_{01}Q_{01}) + \cdots + (P_{t5}/P_{05})(P_{05}Q_{05})}{P_{01} \times Q_{01} + \cdots + P_{05} \times Q_{05}} \times 100 \\ &= \frac{P_{t1} \times Q_{01} + \cdots + P_{t5} \times Q_{05}}{P_{01} \times Q_{01} + \cdots + P_{05} \times Q_{05}} \times 100 \\ &= \frac{\sum_{i=1}^{5} P_{ti}Q_{0i}}{\sum_{i=1}^{5} P_{0i}Q_{0i}} \times 100 \end{aligned} \tag{19.6}$$

Through this derivation, we see that this index is the same index as in Eq. 19.5. That is, this price index is weighted on the basis of the base-year quantities. In other words, the numerator is the value of the expenditures in year t that are necessary to

buy the same quantity of the commodities as was purchased in the base year. This greatly simplifies updating the index, particularly in that most aggregate business indexes contain a large number of items. The denominator is the value of the expenditures required to buy a given amount in the base year. This formula is referred to as the *Laspeyres price index*. The CPI is a Laspeyres index.[1] The disadvantage of this index is that it tends to give more weight to those items that show a dramatic price increase. A sharp increase in a particular commodity's price is typically accompanied by a decrease in the demand (measured by Q) for this item (consumers may be substituting another item). The Laspeyres index fails to adjust for this situation, but even so, its advantages outweigh its disadvantages.

From the data listed in Tables 19.1, 19.2, and 19.3, and from assumptions, we have the price and quantity information listed in Table 19.4.

Substituting data from Table 19.4 into Eq. 19.6, we obtain the price indexes for 1990 and 1991:

$$I_{90} = \frac{180 + 525 + 270 + 770 + 200}{150 + 450 + 220 + 440 + 160} = \frac{1,945}{1,420} = 137$$

$$I_{91} = \frac{225 + 600 + 320 + 990 + 100}{1,420} = \frac{2,235}{1,420} = 157$$

These figures imply that the *weighted aggregative price indexes* estimated from five commodities for 1990 and 1991 are 37 % and 57 % higher than those of 1989, respectively.

19.2.4.2 The Paasche Price Index

The only difference between a *Paasche price index* and a Laspeyres index is that a Paasche index employs the current-year quantities (Q_t) rather than the base-year quantities (Q_0). Formally,

$$I_t = \frac{\sum_{i=1}^{n} P_{ti} Q_{ti}}{\sum_{i=1}^{n} P_{0i} Q_{ti}} \times 100 \tag{19.7}$$

In other words, the numerator determines the amount of money necessary to purchase a given amount of commodities in the current year at current-year prices. Accordingly, the denominator determines the amount of money required to buy the current-year quantities at base-year prices. The gross national product (GNP)

[1] Application of CPI in determining inflation rate and interest rate will be discussed in Application 19.3.

Table 19.4 Price and quantity for five commodities, 1989–1991

	1989		1990		1991	
Commodity	Price	Quantity	Price	Quantity	Price	Quantity
Eggs	1.00	150	1.20	160	1.50	180
Milk	1.50	300	1.75	250	2.00	300
Butter	1.10	200	1.35	180	1.60	250
Bread	0.40	1,100	0.70	1,000	0.90	1,050
Shirts	16.00	10	20.00	15	10.00	20

deflator is a Paasche index. This *GNP deflator* is broader than the CPI because it includes not only consumer goods and services but also investment goods, goods and services purchased by government, and goods and services that enter into world trade.

The complexity of updating the reference-year quantities for a Paasche index makes it difficult (and often impossible) to apply. Furthermore, because it reflects changes in both price and quantity, we cannot use it to reflect price changes between two periods. In addition, it tends to understate price increases and to overstate price decreases because it simultaneously reflects the quantity changes in the demand (Q). Its obvious advantage is that it uses current-year quantities, which provide a realistic and up-to-date estimate of total expense.

Substituting related data from Table 19.4 into Eq. 19.7, we obtain the Paasche indexes for 1990 and 1991:

$$I_{90} = \frac{192 + 437.5 + 243 + 700 + 300}{160 + 375 + 198 + 400 + 240} = \frac{1,872.5}{1,373} = 136$$

$$I_{91} = \frac{270 + 600 + 400 + 945 + 200}{180 + 450 + 275 + 420 + 320} = \frac{2,415}{1,645} = 147$$

19.2.4.3 Fisher's Ideal Price Index

Fisher's ideal price index offers a compromise between the Laspeyres price index and the Paasche price index. This index is found by multiplying the square root of the Laspeyres index by the Paasche price index. Using Eqs. 19.6 and 19.7, we obtain Fisher's ideal price index (FI):

$$\mathrm{FI}_t = \sqrt{\left(\frac{\sum_{i=1}^{n} P_{ti}Q_{ti}}{\sum_{i=1}^{n} P_{0i}Q_{ti}}\right)\left(\frac{\sum_{i=1}^{n} P_{ti}Q_{0i}}{\sum_{i=1}^{n} P_{0i}Q_{0i}}\right)} \tag{19.8}$$

Hence, Fisher's ideal price index lies between the Laspeyres price index and the Paasche price index.

By substituting related information from this section, we find that

$$\text{FI}_{90} = \sqrt{(136)(137)} = 136$$

$$\text{FI}_{91} = \sqrt{(147)(157)} = 152$$

As we expected, the Fisher index is larger than the Paasche index and smaller than the Laspeyres index.

19.3 Quantity Indexes

A *quantity index* measures a change in quantity from a base year to a particular year; such quantities include the volume of industrial production, the physical volume of imports and exports, quantities of goods and services consumed, and the volume of stock transactions. In this section we will discuss two major kinds of quantity indexes – weighted aggregative quantity indexes and weighted relative quantity indexes.

19.3.1 Laspeyres Quantity Index

The Laspeyres quantity index is derived by simply interchanging the P's and Q's in the Laspeyres price index:

$$I_t = \frac{\sum_{i=1}^{n} Q_{ti} P_{0i}}{\sum_{i=1}^{n} Q_{0i} P_{0i}} \times 100 \tag{19.9}$$

This index represents the total cost of the quantities in the year in question at base-year prices as a percentage of the total cost of the base-year quantities. Because prices are kept constant, any change in the index is due to the change in quantities between the base year and the year in question.

From the data in Table 19.5, we can compute the Laspeyres quantity indexes for 1989–1991:

$I_{89} = 57{,}000/57{,}000 \times 100 = 100$
$I_{90} = 74{,}000/57{,}000 \times 100 = 130$
$I_{91} = 91{,}000/57{,}000 \times 100 = 160$

Table 19.5 Worksheet for calculating the Laspeyres quantity index

	1989		1990	1991			
Commodity	Q_{0i}	P_{0i}	Q_{1i}	Q_{2i}	$Q_{0i}P_{0i}$	$Q_{1i}P_{0i}$	$Q_{2i}P_{0i}$
Automobiles	40	1,000	50	60	40,000	50,000	60.000
Computers	30	500	40	50	15,000	20,000	25,000
Televisions	10	200	20	30	2,000	4,000	6,000
					57,000	74,000	91,000

Hence, the indexes for 1990 and 1991 indicate that the cost of the three commodities increased 30 % and 60 %, respectively. The price has been held constant, so the change in the index is due to changes in the quantities of the commodities for the period in question. In other words, the 1990 and 1991 indexes show that the quantities of the goods increased 30 % and 60 %, respectively, from the 1989 base year.

19.3.2 Paasche Quantity Index

The same relationship that exists between the Laspeyres price and quantity indexes also exists between the Paasche price and quantity indexes. Interchanging the P's and Q's in the Paasche price index creates the Paasche quantity index:

$$I_t = \frac{\sum_{i=1}^{n} Q_{ti}P_{ti}}{\sum_{i=1}^{n} Q_{0i}P_{ti}} \times 100 \tag{19.10}$$

In other words, the Paasche quantity index represents the total cost of the purchased quantities in a given year as a percentage of what the base-year quantities would have cost had they been purchased during that year.

From the data in Table 19.6, we can calculate the Paasche quantity indexes for 1989–1991:

$I_{89} = 100{,}500/100{,}500 \times 100 = 100$
$I_{90} = 135{,}000/105{,}000 \times 100 = 129$
$I_{9\mathrm{I}} = 191{,}000/124{,}000 \times 100 = 154$

Prices are held fixed in the equation, so a change between numerator and denominator reflects a change in the quantities between the 2 years. In other words, 1990 and 1990 saw a 23 % and a 54 % increase in quantity, respectively, over the 1989 quantity.

Table 19.6 Worksheet for calculating the Paasche quantity index

	1989		1990		1991						
Commodity	Q_{0i}	P_{0i}	Q_{1i}	P_{1i}	Q_{2i}	P_{2i}	$Q_{0i}P_{0i}$	$Q_{0i}P_{1i}$	$Q_{1i}P_{1i}$	$Q_{0i}P_{2i}$	$Q_{2i}P_{2i}$
Automobiles	40	2,000	50	2,100	60	2,500	80,000	84,000	105,000	100,000	150,000
Computers	30	600	40	600	50	700	18,000	18,000	24,000	21,000	35,000
Televisions	10	250	20	300	20	300	2,500	3,000	6,000	3,000	6,000
							100,500	105,000	135,000	124,000	191,000

19.3.3 Fisher's Ideal Quantity Index

As you may have guessed, there is a compromise between the Laspeyres quantity index and the Paasche quantity index. This compromise is called *Fisher's ideal quantity index (FIQ)*. It is computed as follows:

$$\text{FIQ} = \sqrt{(\text{Laspeyres quantity index})(\text{Paasche quantity index})} \tag{19.11}$$

Like Fisher's price index, Fisher's quantity index lies between the corresponding Laspeyres and Paasche indexes.

Substituting related information into Eq. 19.11 yields

$$\text{FIQ}_{90} = \sqrt{(130)(129)} = 129$$

$$\text{FIQ}_{91} = \sqrt{(160)(154)} = 157$$

19.3.4 FRB Index of Industrial Production

Probably the most widely used and best-known quantity index in the United States is the Federal Reserve Board (FRB) index of industrial production. This index measures changes in the physical volume of output of manufacturing, mining, and utilities. The *FRB index of industrial production* is closely watched by business executives, economists, and financial analysts as a major indicator of the physical output of the economy. It is one of the roughly coincident indicators used by the National Bureau of Economic Research as a cyclical indicator.[2]

The FRB index of industrial production is the weighted arithmetic mean of the quantity relatives as defined as

$$I_t = \frac{\sum_{i=1}^{n} \left(\frac{Q_{ti}}{Q_{0i}} \times 100\right) Q_{0i}P_{0i}}{\sum_{i=1}^{n} Q_{0i}P_{0i}}$$

where Q_{ti}/Q_{0i} is the quantity relative and $Q_{0i}P_{0i}$ is the weight. From the proof of Eq. 19.6, we know this equation is identical to Eq. 19.9.

Numerous problems plague the use of both quantity relative (Q_{ti}/Q_{0i}) and value weights (Q_{0i}/Q_{0i}). Because many industries cannot easily provide physical output data for the quantity relatives, such related data as shipments and employee hours

[2] See Application 19.2 in Sect. 19.6.

worked that tend to move parallel to output are sometimes used instead. Value-added data instead of final product data are also used as weights to avoid the problem of double counting. For example, if the value of the final product were used for a tire company that sells its tires to an automobile company and the value of the final product of the automobile company were also used, the tires that went into making the automobile would be counted twice. A firm's value added is conceptually equivalent to the total of its factor-of-production payments: rent, interest, wages, and profits.

19.4 Value Index

A *value index* measures the total cost of the purchased quantities in a given year at the prices prevailing during that year compared to the cost of the purchased quantities in the base period at base-year prices. The value index is

$$I_t = \frac{\sum_{i=1}^{n} Q_{ti}P_{ti}}{\sum_{i=1}^{n} Q_{0i}P_{0i}}(100) \tag{19.12}$$

Thus, the value index reflects simultaneous changes in quantities and prices for the period in question.

From the data in Table 19.7, we can compute year-to-year changes in the value index:

For 1989–1990,

$$I = (129,000/57,000)(100) = 226$$

For 1990–1991,

$$I = (191,000/129,000)(100) = 148$$

Again, the value index reflects changes from year to year that are due to changes both in prices and in quantities.

19.5 Stock Market Indexes

A *stock market index* is a statistical measure that shows how the prices of a group of stocks change over time.[3] A stock market index encompasses either all or only a portion of stocks in its market. Stock market indexes employ different weighting

[3] Stock index options and futures will be discussed in Appendices 1 and 2, respectively.

Table 19.7 Worksheet for calculating the value index

Commodity	1989 $P_{0i}Q_{0i}$	1990 $P_{1i}Q_{1i}$	1991 $P_{2i}Q_{2i}$
Automobiles	40,000	100,000	150,000
Computers	15,000	24,000	35,000
Televisions	2,000	5,000	6,000
	57,000	129,000	191,000

schemes, so we can use this basis to categorize the indexes by type. The three most common types of stock market indexes are market-value-weighted indexes, price-weighted indexes, and equally weighted indexes. Price per share in current period (P_0), price per share in next period (P_1), number of shares outstanding in current period (Q_0), and number of shares outstanding in next period (Q_1) are listed in Table 19.8. These data are used to illustrate the various weighting schemes and to provide information about the weights applied to the major stock market indexes.

19.5.1 Market-Value-Weighted Index

The *market-value-weighted index* is similar to the value index given in Eq. 19.12 and discussed in Sect. 19.4. We compute the index by taking the ratio of the market value of the outstanding shares at time t to their market value at the initial period. From Table 19.8, we calculate the market-value-weighted index as follows:

$$I = \frac{\sum_{i=1}^{3} Q_{ti}P_{ti}}{\sum_{i=1}^{3} Q_{0i}P_{0i}} = \frac{55(100) + 100(60) + 300(40)}{60(100) + 90(50) + 250(20)}(100)$$
$$= \frac{23,500}{15,500}(100) = 152$$

This figure implies that the market-value-weigh ted index increased by 52 % from the base period to the current period.

Standard & Poor's 500 Composite Index is an example of a market-value-weighted index. The *S&P 500 index* comprises industrial firms, utilities, transportation firms, and financial firms. Changes in the index are based on changes in the firms' total market value with respect to a base year. Currently, the base period ($1941 - 1943 = 10$) for the S&P 500 index is stated formally as follows:

Table 19.8 Price per share and outstanding shares for stocks, A, B, and C

Stock	P_0	P_1	Q_0	Q_1
A	100	100	60	55
B	50	60	90	100
C	20	40	250	300

$$\text{S\&P 500 index} = \frac{\sum_{i=1}^{500} P_{ti}Q_{ti}}{\sum_{i=1}^{500} P_{0i}Q_{0i}}(10) \tag{19.13}$$

where

P_{0i} = per share stock price at base year 0
P_{ti} = per share stock price at index data t
Q_{0i} = number of shares for firm i at base year 0
Q_{ti} = number of shares for firm i at index year t

The index is multiplied by an index set equal to 10. The specification of this index is identical to that of the value index indicated in Eq. 19.12. The fluctuation of the S&P 500 index during 1980–2011 is presented in Fig. 19.1.

The New York Stock Exchange (NYSE) also publishes a market index, which differs in only two respects from the S&P 500 index. First, the NYSE index includes the stocks of *all* firms listed on the NYSE, whereas the S&P 500 index includes only a portion of the firms on the exchange. In addition, the NYSE index uses a base index of 50 (as opposed to 10), which was chosen to represent an approximate price of an average share in December 1965.

19.5.2 Price-Weighted Index

The *price-weighted index* shows the change in the average price of the stocks that are included in the index. Using the data from Table 19.8, we can compute the price-weighted index as follows:

$$\begin{aligned} I &= \frac{(100+60+40)/3}{(100+50+20)/3}(100) \\ &= \frac{200/3}{170/3}(100) \\ &= 117.65 \end{aligned}$$

The closest thing to a true price-weighted stock market index is the *Dow Jones Industrial Average* (DJIA). Simply stated, the DJIA is an arithmetic average of the stock prices that make up the index. The DJIA originally assumed a single share of

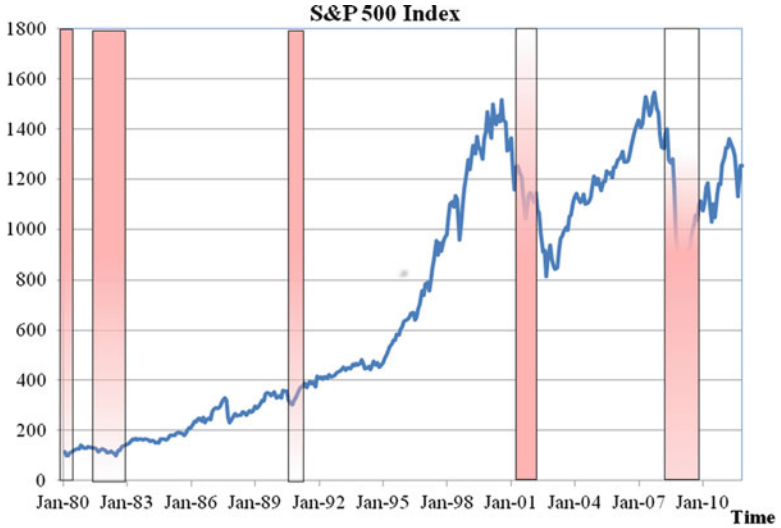


Fig. 19.1 The S&P 500, January 1980–November 2011 (*monthly averages*: 1941–1943 = 10; shaded areas represent periods of business recessions) (*Source*: National Bureau of Economic Research http://www.nber.org/ and Yahoo Finance)

each stock in the index, and the total of the stock prices was divided by the number of stocks that made up the index[4]:

$$\text{DJIA}_t = \frac{\frac{\sum_{i=1}^{30} P_{ti}}{30}}{\frac{\sum_{i=1}^{30} P_{0i}}{30}} \tag{19.14}$$

Today the index is adjusted for stock splits and the issuance of stock dividends:

$$\text{DJIA}_t = \frac{\sum_{i=1}^{30} \frac{P_{ti}}{\text{AD}_t}}{\sum_{i=1}^{30} P_{0i}} \tag{19.14a}$$

where P_{it} = the closing price of stock i on day t and AD = the adjusted divisor on day t. This index is similar to the simple price index given in Eq. 19.1 except for the stock splits adjusted over time. The adjustment process is illustrated in Table 19.9.

[4] There are 30 blue-chip firms included in this index. Their names appear in Table 9.1.

Table 19.9 Adjustment of DJIA divisor to allow for a stock split

Before split		After 2-for-l stock split by stock A
Stock	Price	Price
A	60	30
B	30	30
C	20	20
D	10	10
	120	90

$$\text{Average before split} = \frac{120}{4} = 30$$

$$\text{Adjusted divisor} = \frac{\text{Sum of prices after the split}}{\text{Average before split}} = \frac{90}{30} = 3$$

$$\text{Average after split} = \frac{90}{3} = 30$$

$$\text{Before} - \text{split divisor} = 4 \quad \text{After} - \text{split divisor} = 3$$

Alternatively, the average after split can be calculated as

$$\text{Average} = \frac{30 \times 2 + 30 + 20 + 10}{4} = 30$$

This average is identical to that obtained by using the adjusted-divisor approach.

As Table 19.9 shows, the adjustment process is designed to keep the index value the same as it would have been if the split had not occurred. Similar adjustments have been made when it has been found necessary to replace one of the component stocks with the stock of another company, thus preserving the consistency and comparability of index values at different points in time.

Nevertheless, the adjustment process used for the DJIA has its share of critics. Because price weighting itself causes high-priced stocks to dominate the series, the same effect can cause a shift in this balance when rapidly growing firms split their stock. For example, a 20 % increase in the price of stock A in Table 19.9 would in itself have caused a 10 % increase in the value of the sample index before the split, whereas a 20 % increase in the price of stock B would have caused only a 5 % increase in the index value. After the 2-for-1 split of stock A, a 20 % increase in either stock A or stock B would have the same effect on the index value (a 6.7 % increase); a downward shift in the importance of stock A relative to that of the other stocks in the sample has occurred. This effect could relegate the stock of the fastest-growing companies to a position of *least* importance in determining index values.

19.5.3 Equally Weighted Index

The *equally weighted index* is based on the supposition that an equal amount is invested in each of the stocks included in the index. Hence, in computing the index, we will assume that $1,000 is invested in each of the three stocks. Using the price

information listed in Table 19.8, we find the following numbers of shares purchased in the initial period.

Stock	Number of shares at initial period
A	10(1,000/100)
B	20(1,000/50)
C	50(1,000/20)

Using this information and the price information listed in Table 19.8, we can calculate the equally weighted index:

$$I = \frac{10(100) + 20(60) + 50(40)}{10(100) + 20(50) + 50(20)}(100)$$
$$= \frac{4,200}{3,000}(100) = 140$$

The equally weighted index is based on the changes in the price of the individual stocks, given that an equal amount of money is initially invested in each stock. In other words, the index keeps the number of shares constant while providing for changes in the pre-share price. This is similar to the Laspeyres price index (Eq. 19.6), except that the initial quantity should be determined by the equal-amount-net-investment assumption. One of the two Wilshire 5000 equity indexes is an equally weighted index. The market-value-weighted Wilshire 5000 equity index is discussed in the next section.

19.5.4 Wilshire 5000 Equity Index

The Wilshire 5000 equity index, which includes 5,000 stocks, is compiled by both market-value-weighted and equally weighted approaches. The market-value-weighted approach is identical to the value index given in Eq. 19.12. The equally weighted approach is identical to that discussed in the last section. This index is being used increasingly, because it contains most equity securities available for investment, including all NYSE and AMEX issues and the most active stocks traded on the over-the-counter (OTC) market.

The following formula is used to compute the market-value-weighted Wilshire 5000 equity index:

$$I_t = I_{t-1}\left[\frac{\sum_{j=1}^{N}(S_{jt})P_{jt}}{\sum_{j=1}^{N}(S_{jt-1})P_{jt-1}}\right] \tag{19.15}$$

Table 19.10 Value-weighted Wilshire 5000 equity index

Date	Wilshire index[a]	Date	Wilshire index[a]
1/31/1989	2,917.261	8/31/1990	3,053.601
2/28/1989	2,857.863	9/28/1990	2,879.335
3/31/1989	2,915.072	10/31/1990	2,833.986
4/28/1989	3,053.132	11/30/1990	3,015.022
5/31/1989	3,162.609	12/31/1990	3,101.355
6/30/1989	3,137.008	1/31/1991	3,245.346
7/31/1989	3,377.403	2/28/1991	3,484.851
8/31/1989	3,440.843	3/28/1991	3,583.671
9/29/1989	3,426.656	4/30/1991	3,587.924
10/31/1989	3,320.354	5/31/1991	3,719.297
11/30/1989	3,367.637	6/28/1991	3,545.470
12/29/1989	3,419.879	7/31/1991	3,705.893
1/31/1990	3,163.301	8/30/1991	3,795.043
2/28/1990	3,201.205	9/30/1991	3,743.976
3/30/1990	3,273.458	10/31/1991	3,807.081
4/30/1990	3,172.327	11/29/1991	3,649.992
5/31/1990	3,448.484	12/31/1991	4,041.102
6/29/1990	3,424.366	01/31/1992	4,027.770
7/31/1990	3,384.365		

[a]2/30/1980 base = 1,404.596

where

I_t = index value for the tth period
N = number of stocks in the index
P_{jt} = price of the jth security for the tth period
S_{jt} = shares outstanding of the jth security for the tth period
P_{jt-1} = price of the jth security for the $(t-1)$th period
S_{jt-1} = shares outstanding of the jth security for the $(t-1)$th period

In the event that P_{jt} is not available for a given security, that security is dropped from the summations. If P_{jt-1} is not available but P_{jt} is – that is, if a security has just resumed trading – the last available price is substituted for P_{jt-1}.

As an example, we present in Table 19.10 monthly equity values for the Wilshire 5000 equity index from January 1989 to January 1992. In this table, only the value-weighted index appears. Figure 19.2 is a graph of the Wilshire 5000 equity index from January 1970 to December 1991. The monthly Wilshire equity index (Table 19.10) can be used to calculate the market rates of return by the method discussed in Appendix 2 in Chap. 2. Monthly rates of return calculated from the data of Table 19.10 in terms of the value-weighted index are presented in Table 19.11. In Table 19.11, price appreciation represents the percentage change of index, and the total return is equal to price appreciation plus the dividend yield. Note that the Wilshire index can be used in place of the S&P 500 as an NBER leading economic indicator, a topic discussed in Sect. 19.6 of this chapter.

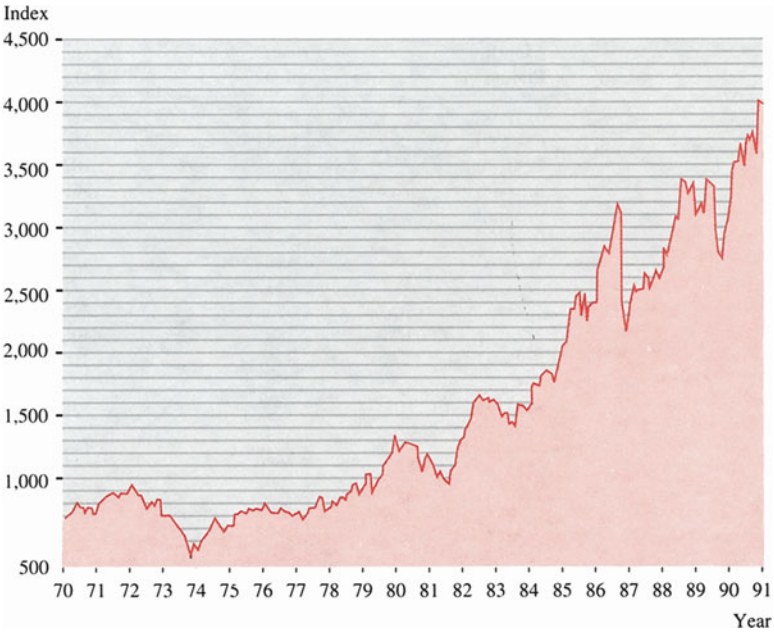


Fig. 19.2 The Wilshire 5000, January 1971–October 1990 (*Source*: Wilshire 5000 Equity Index, Wilshire 5000® a registered Service Mark of Wilshire Associates Incorporated, Santa Monica, California)

19.6 Business and Economic Applications

Application 19.1 Deflation of Value Series by Price Indexes. Corporate executives and economists are often interested in dividing time series data expressed in monetary terms into two components – a "real," or quantity, component and a price component. Together, these two components express a value series in terms of price × quantity. Hence, if we are interested only in the quantity component of a time series, we can find it simply by dividing the value series by price. Often we use a price index as a deflator. With the available information, it is important to use an index that is relevant to the time series in question. For example, one might use the consumer price index (CPI) to calculate real weekly wage and then use the GDP deflator to calculate real GDP.

If we want to calculate the real weekly wage for 2007–2009, we can use the consumer price index to deflate the nominal weekly wage and obtain the real wage rate as indicated in Table 19.12.

Table 19.13 shows nominal GDP and the implicit price deflator for GDP and real GDP for the period 2000–2009. Figure 19.3 plots both nominal and real GDP data. Calculating real GDP gives insight into the "real" growth in the economy. We do not want to be misled by the trend of GDP when prices generally increase at a rate

Table 19.11 Monthly returns for value-weighted Wilshire 5000 equity index

Month ending	Price appreciation (%)[a]	Dividends yield (%)	Total return (%)
1/31/1989	6.531	0.281	6.812
2/28/1989	−2.036	0.368	−1.668
3/31/1989	2.002	0.272	2.274
4/28/1989	4.736	0.180	4.916
5/31/1989	3.586	0.475	4.061
6/30/1989	−0.810	0.233	−0.577
7/31/1989	7.663	0.213	7.876
8/31/1989	1.878	0.394	2.272
9/29/1989	−0.412	0.239	−0.173
10/31/1989	−3.102	0.182	−2.920
11/30/1989	1.424	0.342	1.766
12/29/1989	1.551	0.265	1.816
1/31/1990	−7.503	0.163	−7.340
2/28/1990	1.198	0.390	1.588
3/30/1990	2.257	0.242	2.499
4/30/1990	−3.090	0.210	−2.880
5/31/1990	8.705	0.426	9.131
6/29/1990	−0.699	0.216	−0.483
7/31/1990	−1.168	0.202	−0.966
8/31/1990	−9.773	0.363	−9.410
9/28/1990	−5.707	0.213	−5.494
10/31/1990	−1.575	0.235	−1.340
11/30/1990	6.388	0.429	6.817
12/31/1990	2.863	0.311	3.1 74
1/31/1991	4.643	0.215	4.858
2/28/1991	7.380	0.400	7.780
3/28/1991	2.836	0.210	3.046
4/30/1991	0.119	0.198	0.317
5/31/1991	3.662	0.348	4.010
6/28/1991	−4.674	0.209	−4.465
7/31/1991	4.525	0.171	4.696
8/30/1991	2.406	0.356	2.762
9/30/1991	−1.346	0.198	−1.148
10/31/1991	1.686	0.151	1.837
11/29/1991	−4.126	0.307	−3.819
12/31/1991	10.715	0.265	10.980
01/31/1992	−0.330	0.132	−0.198

[a]Represents monthly percentage change in Wilshire 5000 equity index

greater than zero. A price component (price deflator) provides information about changes in the nominal component of output. Formally, we compute real GDP as follows:

$$\text{Real GDP} = \frac{\text{Nominal GDP}}{\text{GDP deflator}/100} \tag{19.16}$$

Table 19.12 Calculation of real weekly wages, 2007–2009

(1) Year	(2) Average weekly wage ($)	(3) Consumer price index (1982 − 1984 = 100)	(4) Real weekly wage ($)
2007	590.04	207.3	284.58
2008	607.99	215.3	282.46
2009	616.37	214.5	287.29

Source: Economic Report of the President, January 2010

Table 19.13 Nominal GDP, GDP deflators, and real GDP (2000–2009)

Year	Nominal GDP	GDP deflators	Real GDP
2000	9,951.51	88.647	11,226.00
2001	10,286.24	90.650	11,347.20
2002	10,642.39	92.118	11,553.00
2003	11,142.10	94.100	11,840.70
2004	11,867.68	96.770	12,263.80
2005	12,638.40	100.000	12,638.40
2006	13,398.83	103.257	12,976.20
2007	14,077.71	106.214	13,254.10
2008	14,441.47	108.483	13,312.20
2009	14,258.61	109.777	12,988.70

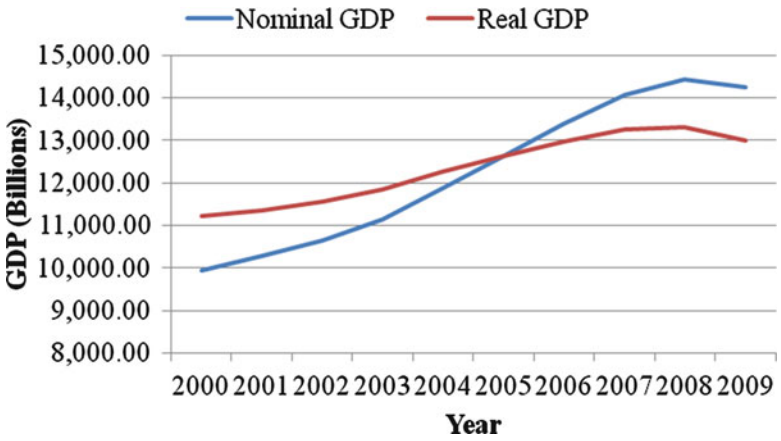


Fig. 19.3 Nominal versus real GDP, 2000–2009

Equation 19.16 is used to calculate real GDP for the period 2000–2009; the results are presented in the last column of Table 19.13.

To calculate the growth rate for nominal GDP (X_t), GDP deflator (Y_t), and real GDP (Z_t) from the data listed in Table 19.13, we run the following regressions:

Regression Analysis: log(nominal GDP) versus Year

```
The regression equation is
log(nominal GDP) = -83.2 + 0.0462 Year

Predictor       Coef   SE Coef       T      P
Constant    -83.248     5.613  -14.83  0.000
Year       0.046223  0.002800   16.51  0.000

S = 0.0254361   R-Sq = 97.1%   R-Sq(adj) = 96.8%

Analysis of Variance

Source           DF       SS       MS       F      P
Regression        1  0.17627  0.17627  272.44  0.000
Residual Error    8  0.00518  0.00065
Total             9  0.18144

           log(nominal
Obs  Year       GDP)       Fit   SE Fit  Residual   St Resid
  1  2000    9.20548  9.19794  0.01495   0.00754       0.37
  2  2001    9.23856  9.24416  0.01268  -0.00560      -0.25
  3  2002    9.27260  9.29039  0.01066  -0.01779      -0.77
  4  2003    9.31849  9.33661  0.00907  -0.01812      -0.76
  5  2004    9.38157  9.38283  0.00816  -0.00126      -0.05
  6  2005    9.44450  9.42906  0.00816   0.01544       0.64
  7  2006    9.50292  9.47528  0.00907   0.02764       1.16
  8  2007    9.55235  9.52150  0.01066   0.03085       1.34
  9  2008    9.57786  9.56772  0.01268   0.01014       0.46
 10  2009    9.56512  9.61395  0.01495  -0.04883      -2.37R

R denotes an observation with a large standardized residual.
Durbin-Watson statistic = 0.956295
```

Fig. 19.4 MINITAB output of Eq. 19.17a

$$\log_e X_t = a_0 + a_1 t + e_{1t} \tag{19.17a}$$

$$\log_e Y_t = b_0 + b_1 t + e_{2t} \tag{19.17b}$$

$$\log_e Z_t = c_0 + c_1 t + e_{3t} \tag{19.17c}$$

The justification for using these equations to estimate growth rate appears in Eq. 18.13 of the last chapter. MINITAB results of Eqs. 19.17a, 19.17b, and 19.17c are presented in Figs. 19.4, 19.5, and 19.6, respectively.

From Figs. 19.4, 19.5, and 19.6, we obtain the following estimated slopes:

$$\underset{(34.27)}{\hat{a}_1 = 561} \quad \underset{(0.0922)}{\hat{b}_1 = 2.52} \quad \underset{(26.83)}{\hat{c}_1 = 254}$$

Regression Analysis: log(GDP Deflators) versus Year

```
The regression equation is
log( GDP Deflators) = -46.5 + 0.0255 Year

Predictor       Coef    SE Coef       T      P
Constant     -46.487      1.699  -27.35  0.000
Year       0.0254822  0.0008478   30.06  0.000

S = 0.00770053   R-Sq = 99.1%   R-Sq(adj) = 99.0%

Analysis of Variance

Source          DF        SS        MS       F      P
Regression       1  0.053571  0.053571  903.41  0.000
Residual Error   8  0.000474  0.000059
Total            9  0.054045

             log(GDP
Obs  Year  Deflators)     Fit   SE Fit  Residual  St Resid
  1  2000     4.48466  4.47776  0.00453   0.00690      1.11
  2  2001     4.50701  4.50324  0.00384   0.00376      0.56
  3  2002     4.52307  4.52873  0.00323  -0.00566     -0.81
  4  2003     4.54436  4.55421  0.00275  -0.00985     -1.37
  5  2004     4.57234  4.57969  0.00247  -0.00735     -1.01
  6  2005     4.60517  4.60517  0.00247  -0.00000     -0.00
  7  2006     4.63722  4.63066  0.00275   0.00657      0.91
  8  2007     4.66546  4.65614  0.00323   0.00932      1.33
  9  2008     4.68659  4.68162  0.00384   0.00497      0.74
 10  2009     4.69845  4.70710  0.00453  -0.00865     -1.39

Durbin-Watson statistic = 0.909893
```

Fig. 19.5 MINITAB output of Eq. 19.17b

The standard errors are listed under the estimates. Note that $\hat{a_1}$, $\hat{b}1$, and $\hat{c_1}$, are growth rate estimates.

From these estimates, we know that the growth rate for nominal GDP is approximately equal to the growth rate of the GDP deflator plus the growth rate of real GDP.

Application 19.2 Using Business and Economic Index Numbers to Predict Business Cycles. The National Bureau of Economic Analysis publishes a series of 26 categories of economic data reflecting movements in the business cycle (see Table 19.14). These are called *indicators* of economic activity. Of the 26 indicators, 12 make up an index of leading indicators. A leading indicator's "lead" must be greater than 3 months, and a lagging indicator follows economic activity by more than 3 months. Indicators of economic activity that lead or lag economic activity by 3 months or less are called coincident indicators.

The 12 components of the index of leading indicators are:

Regression Analysis: log(GDP Deflators) versus Year

```
The regression equation is
log(Real GDP) = -32.2 + 0.0207 Year

Predictor       Coef   SE Coef      T      P
Constant    -32.156     4.364  -7.37  0.000
Year       0.020741  0.002177   9.53  0.000

S = 0.0197743   R-Sq = 91.9%   R-Sq(adj) = 90.9%

Analysis of Variance

Source          DF        SS        MS      F      P
Regression       1  0.035490  0.035490  90.76  0.000
Residual Error   8  0.003128  0.000391
Total            9  0.038618

           log(Real
Obs  Year      GDP)      Fit   SE Fit  Residual  St Resid
  1  2000   9.32599  9.32535  0.01162   0.00064      0.04
  2  2001   9.33673  9.34609  0.00986  -0.00936     -0.55
  3  2002   9.35470  9.36683  0.00829  -0.01213     -0.68
  4  2003   9.37930  9.38757  0.00705  -0.00827     -0.45
  5  2004   9.41441  9.40831  0.00635   0.00610      0.33
  6  2005   9.44450  9.42905  0.00635   0.01544      0.82
  7  2006   9.47087  9.44979  0.00705   0.02108      1.14
  8  2007   9.49206  9.47053  0.00829   0.02153      1.20
  9  2008   9.49644  9.49127  0.00986   0.00516      0.30
 10  2009   9.47184  9.51202  0.01162  -0.04018     -2.51R

R denotes an observation with a large standardized residual.
Durbin-Watson statistic = 0.886178
```

Fig. 19.6 MINITAB output of Eq. 19.17c

1. Average weekly hours of manufacturing workers
2. Average weekly initial claims for unemployment insurance
3. The real value of manufacturers' new orders for consumer goods and materials
4. The index of new business formations
5. The index of 500 common stock prices
6. Contracts and orders for plant and equipment in 1972 dollars
7. The index of new private housing starts authorized by local building permits
8. The ratio of price to unit labor cost, manufacturing

Table 19.14 Cyclical indicators: short list of the National Bureau of Economic Research

Leading indicators
Average hourly workweek, production workers, manufacturing
Average weekly initial claims, state unemployment insurance
Index of net business formation
New orders, durable-goods industries
Contracts and orders, plant and equipment
Index of new building permits, private housing units
Change in book value, manufacturing, and trade inventories
Index of industrial materials prices
Index of stock prices, 500 common stocks
Corporate profits after taxes (quarterly)
Index: ratio of price to unit labor cost, manufacturing
Change in consumer installment debt
Roughly coincident indicators
GNP in current dollars
GNP in 1958 dollars
Index of industrial production
Personal income
Manufacturing and trade sales
Sales of retail stores
Employees on nonagricultural payrolls
Unemployment rate, total
Lagging indicators
Unemployment rate, persons unemployed 15 weeks or over
Business expenditures, new plant, and equipment
Book value, manufacturing, and trade inventories
Index of labor cost per unit of output in manufacturing
Commercial and industrial loans outstanding in large commercial banks
Banks rates on short-term business loans

Source: U.S. Department of Commerce

9. The net change in inventories on hand and on order in 1972 dollars
10. The change in sensitive materials prices
11. The change in total consumer and business credit outstanding
12. Corporate profits after taxes (quarterly)

Table 19.15 presents information on the dates and durations of business cycles as determined by the National Bureau of Economic Research (NBER). It is particularly interesting to compare the dates and durations of business cycles to movements in the stock market. Comparing the graphs of the S&P 500 (Fig. 19.1) and the Wilshire 5000 stock indexes (Fig. 19.2) to the business cycle information reported in Table 19.15 reveals the relationship between the business cycle and stock market movements. For example, from February 1961 to December 1969, the United States sustained 106 months of economic growth (largely due to government expenditures on the Vietnam War). Figure 19.1 shows that over the

Table 19.15 NBER business cycle reference dates and durations

Trough	Peak	Contractions	Expansions	Trough to trough	Peak to peak
December 1854	June 1857	NA	30	NA	NA
December 1858	October 1860	18	22	48	40
June 1861	April 1865	8	46	30	54
December 1867	June 1869	32	18	78	50
December 1870	October 1873	18	34	36	52
March 1879	March 1882	65	36	99	101
May 1885	March 1887	38	22	74	60
April 1888	July 1890	13	27	35	40
May 1891	January 1893	10	20	37	30
June 1894	December 1895	17	18	37	35
June 1897	June 1899	18	24	36	42
December 1900	September 1902	18	21	42	39
August 1904	May 1907	23	33	44	56
June 1908	January 1910	13	19	46	32
January 1912	January 1913	24	12	43	36
December 1914	August 1918	23	44	35	67
March 1919	January 1920	7	10	51	17
July 1921	May 1923	18	22	28	40
July 1924	October 1926	14	27	36	41
November 1927	August 1929	13	21	40	34
March 1933	May 1937	43	50	64	93
June 1938	February 1945	13	80	63	93
October 1945	November 1948	8	37	88	45
October 1949	July 1953	11	45	48	56
May 1954	August 1957	10	39	55	49
April 1958	April 1960	8	24	47	32
February 1961	December 1969	10	106	34	116
November 1970	November 1973	11	36	117	47
March 1975	January 1980	16	58	52	74
July 1980	July 1981	6	12	64	18
November 1982	July 1990	16	80[a]	28	96[a]
March 1991	?	NA	NA	89[a]	NA

[a]The 80-month duration of the last expansion, the 96-month duration of the last peak-to-peak cycle, and the 89-month duration of the last trough-to-trough cycle are conservative estimates. They assume a peak in July 1989 and, for the last of these, a trough 9 months later. Wartime expansions and cycles containing wartime expansions are *underlined*

same period, the S&P 500 index moved higher, indicating a direct link between economic growth and stock market movements. That is, when the economy is growing, stock prices tend to increase. When the economy is in recession, stock prices decline. In general, stock price movements *lead* changes in economic condition. Therefore, the stock market is one of the leading indicators of the business cycle.

In the business section of the *Home News* on June 1, 1991, a headline read, "Leading Indicators Up Again." In support of this headline, the newspaper

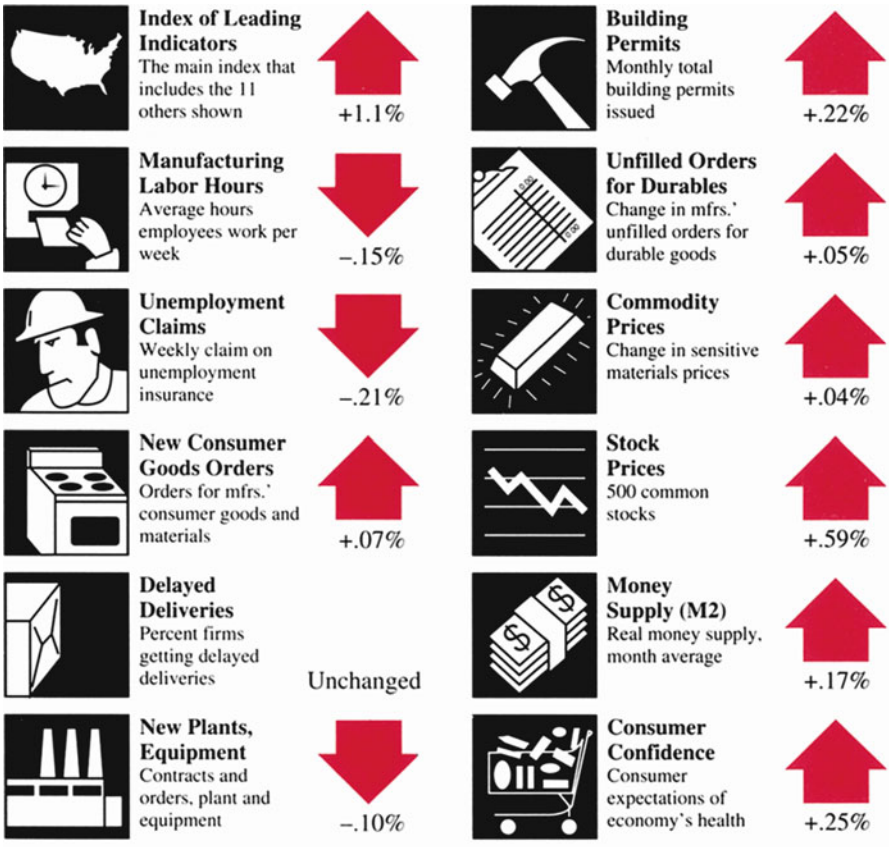


Figure 19.7 Statistics that forecast the economy (*Source*: *Home News*, March 30, 1991. Reprinted by permission of Knight-Ridder Tribune News)

identified the 12 statistics listed in Fig. 19.7 as "statistics that forecast the economy." Of these 12 statistics, 8 increased, 1 experienced no change, and 3 decreased. This example essentially shows that the cyclical indicators in Table 19.14 can be used to forecast economic activities.

Application 19.3 CP1, Inflation Rate, and Interest Rate. The consumer price index (CPI) defined in Sect. 19.2 of this chapter can be used to calculate the inflation rate. For example, using the CPI index presented in Table 2.1 of Chap. 2, we can calculate the inflation for 2010 (I_{2010}) as

$$I_{2010} = \ln\left(\frac{CPI_{2010}}{CPI_{2009}}\right) = \ln\left(\frac{218.1}{214.5}\right) = 1.66\%$$

To try to protect themselves from a loss of purchasing power, investors will demand a return that reflects inflation expectations. This return is called the nominal risk-free interest rate; it represents the observed or published return on a risk-free asset.

The interest rate on 1-year or 3-month Treasury bill is generally used by many financial analysts to approximate the nominal risk-free rate.[5] The 3-month Treasury bill rate can be found in Table 2.1.

The relationship between the nominal risk-free rate and expected inflation can be defined as

$$\begin{aligned}\text{Nominal risk} - \text{free interest rate} = {} & (1 + \text{real risk} - \text{free rate}) \\ & \times (1 + \text{expected inflation rate}) - 1\end{aligned}$$

This equation is known as the Fisher effect. For example, the real risk-free interest rate of 4 % and lender expectations of 6 % inflation over the next year produce a nominal risk-free rate of

$$(1 + 0.04)(1 + 0.06) - 1 = 10.24\%$$

To summarize, the CPI index can be used to calculate the inflation rate, and this can be used to determine the nominal risk-free rate. It is well known that the nominal risk-free rate (Treasury bill rate) is generally used to determine mortgage rate or other lending rate.

19.7 Summary

In this chapter, we examined different index numbers. Price, quantity, and value indexes were explored. Then stock market indexes and applications of index numbers to calculate real wage income and real GNP were discussed in detail. Finally, we demonstrated how the NBER economic indicators are used to forecast business cycles.

Questions and Problems

1. Find the following sales indexes for the accompanying retail sales volume (in thousands of dollars) with the base year indicated.

Year	Sates ($)	Year	Sales ($)
1984	85,390	1987	92,289
1985	86,745	1988	97,725
1986	88,452		
(a) 1984			
(b) 1988			

[5] See Lee, C.F., et al.: *Foundation of Financial Management*. West Publishing Company, Minneapolis/St. Paul (1997), Chap. 3 for detail.

2. An alloy is made up of 27 % metal A, 34 % metal B, and 39 % metal C. During the past year, prices of the metals have changed as shown in the table. Find the simple aggregative index for the cost of the alloy.

Price per pound		
Metal	1988	1989
A	2.08	3.48
B	7.61	7.78
C	4.49	3.80

3. Find the simple aggregative index number for the following data.

	1972		1989	
Item	Unit price	Unit sold	Unit price	Unit sold
A	$2.50	400	$4.00	650
B	16.00	150	19.50	175
C	9.50	250	16.00	350

(a) Find the Laspeyres index number.
(b) Find the Paasche index number.
(c) Find the ideal index number.
(d) Find a physical volume index for 1989, weighting quantities with 1972 base-year prices.

4. Use the data in question 2.

(a) Find the weighted aggregative index for the cost of the alloy.
(b) How can you explain the values of these index numbers in view of the fact that only 2 out of 3 constituent metal prices rose?

5. Say you earned $10,725 in 1975 when the CPI in your city was 150 and earned $29,500 in 1989 when your city's CPI hit 358. Express your 1989 purchasing power as a percentage of your 1975 purchasing power.
6. Assume the CPI for your city had the following values.

Year	1982	1983	1984	1985	1986
Value	210	238	265	300	335

(a) Find the purchasing power of the dollar in each of these years as a proportion of the 1986 dollar.
(b) Explain the meaning of these figures (*Hint*: What percentage of 1986 goods could you buy in these years?)

7. Suppose you have a stock market indicator made up of five common stocks selling at the prices per share indicated in the following table.

Stock	Price per share
A	$106.00
B	87.75
C	49.50
D	32.75
E	23.50

(a) Find the market average.
(b) Suppose stock A split 2 for 1 and stock D split 3 for 1. Find the new denominator for your average.
(c) After the splits, the prices settled to the values shown in the following table. Find the indicator's new value.

Stock	Price per share
A	$54.00
B	90.75
C	53.25
D	11.75
E	23.25

8. For the accompanying data for the retail price of selected appliances, find the Laspeyres retail price index for each year, using 1967 as the base.

	Average unit price			Thousands of units sold
Appliance	1967	1978	1991	1967
A	255	268	310	5,930
B	310	327	323	1,950
C	223	250	265	2,010
D	37	39	42	890

9. Which of the following index numbers could be found by using the data in question 8?

(a) Simple aggregative index
(b) Laspeyres index
(c) Fisher's ideal index
(d) Weighted aggregative index
(e) Paasche index

10. What is an index number? Give some examples of index numbers. Why are they useful?
11. What is the difference between an index constructed with simple averages and an index constructed with weighted averages? If we construct an index of stock

prices in which all the companies are small and approximately the same size, is there any difference between a weighted-average index and a simple-average index? What if we construct an index using companies of many different sizes?

12. What is the difference between the ways the Paasche and Laspeyres price indexes are computed?
13. Under what conditions would you expect Paasche and Laspeyres indexes to be significantly different?
14. What is Fisher's ideal price index? Why might it be better to use than a Paasche or Laspeyres index?
15. One commonly reported index in business is the index of leading economic indicators. What is the purpose of this index? If you were asked to construct your own index of leading economic indicators, what information would you use? How would you construct it?
16. Explain how a price-weighted stock index must be adjusted to reflect the stock split of a company.
17. Some people argue that price indexes do not reflect the improvements in quality of the products we buy. Would this limitation cause estimates of inflation to be too high or too low?
18. What is a base year? What is the value of the index in the base year?
19. What is the consumer price index? What does it measure? How is it constructed?
20. Use the data in Tables 19.5 and 19.6 to compute the Laspeyres index and compare the result to the Paasche index. Also compute Fisher's ideal quantity index.
21. What is a real component? What is a price component? How are the two related?
22. What is the Wilshire 5000 equity index? What securities are included in this index? How is it computed?
23. You are given the following information on prices and quantities for widgets.

Year	Price	Quantity
1985	$1.00	1,000
1986	1.25	1,800
1987	1.31	2,000
1988	1.44	2,222
1989	1.51	2,325
1990	1.85	3,100

Using 1987 as the base year, compute the Laspeyres price index for widgets.

24. Drawing on the data given in question 23, compute the Paasche price index, again using 1987 as the base year.
25. Using your calculations in questions 23 and 24, compute Fisher's ideal price index.

26. On the Island of Crusoe, there are only two goods: coconuts and fish. Suppose you have the following information on the prices and quantities of fish and coconuts.

	Fish		Coconuts	
Year	Q	P	Q	P
1987	100	$3.00	75	$1.00
1988	110	2.90	80	1.02
1989	99	3.12	79	1.05
1990	121	3.45	88	1.15

Using 1988 as the base year, compute the Paasche price index.

27. Using the data from question 26, compute the Laspeyres price index.
28. Using the data from questions 26 and 27, compute Fisher's ideal price index.
29. Using the Laspeyres, Paasche, and Fisher's indexes you computed in questions 26–28, compute the percentage change in price for each year (inflation rate), and compare the results for the three indexes. Which one gives the highest rate? Which one the lowest?
30. Three commonly reported measures of price level are the consumer price index, the producer price index, and the GNP deflator. Explain why these measures may yield different inflation rates.
31. Here are some price indexes for four different types of collectibles.

Collectible	1980	1985	1990
Baseball cards	100	210	275
Paintings	100	195	325
Jewelry	100	250	245
Gold coins	100	199	201

(a) What is the base year?
(b) Which of the collectibles increased the most in price from 1980 to 1990?
(c) Which of the collectibles increased the most in price from 1985 to 1990?
(d) Which of the collectibles increased the most in price from 1980 to 1985?

32. Consider the following price and market-value information for five stocks in 1990.

Stock	Price ($)	Market value ($)
A	100	1,000,000
B	50	3,000,000
C	72	500,000
D	35	1,250,000
E	27	300,000

Compute an equally weighted price average and a value-weighted price average. Explain why the two indexes differ.

33. Now suppose you have the following price and market-value information for the same stocks of question 32 in 1991.

Stock	Price ($)	Market value ($)
A	105	1,500,000
B	30	2,000,000
C	72	800,000
D	25	1,850,000
E	57	900,000

Compute an equally weighted and a value-weighted relative price index.

34. Suppose you are given the following information about wages and prices for 5 years.

Year	Average hourly wage ($)	Consumer price index
1986	8.22	100
1987	9.37	110
1988	10.01	108
1989	11.27	135
1990	15.43	200

(a) Compute real wages for all years between 1986and 1987.
(b) Are workers any better off in 1987 than they were in 1986?

35. Use the data given in question 34 to compute the change in real wages between 1986 and 1990. Are workers any better off?
36. You are given the following cost indexes for three categories of consumer expenditures.

Year	Housing	Food	Transportation
1980	127.2	129.3	151.2
1985	145.6	141.2	170.6
1990	166.7	171.2	200.3

Compute the percentage change in housing, food, and transportation costs between 1980 and 1985. Which expenditure underwent the greatest price increase?

37. Repeat question 36 for the period 1985–1990 and the period 1980–1990.
38. Agricultural economists have data on the price per bushel of wheat, corn, barley, and hops for the last 20 years. They want to obtain a measure for the aggregative price movements of grain over this period. In deciding how to produce this price index, what factors should they take into consideration?

39. The following table shows the price of hamburger over the last 10 years.

Year	Price per pound ($)
1	1.82
2	1.95
3	1.86
4	2.10
5	2.45
6	3.10
7	2.60
8	2.45
9	2.75
10	2.58

(a) Form a price index, using year 1 as the base.
(b) Form a price index, using year 6 as the base.

40. The following table shows the price of gold over the last 10 weeks.

Week	Price per ounce ($)
1	$410.82
2	401.95
3	391.56
4	382.10
5	392.45
6	403.10
7	412.60
8	392.45
9	399.75
10	402.58

(a) Form a price index, using week 1 as the base.
(b) Form a price index, using week 5 as the base.

41. The following table shows the average price, taken monthly, of Widget Company stock over the last year.

Month	Price	Month	Price
1	$41\frac{3}{8}$	7	$50\frac{1}{4}$
2	$42\frac{1}{4}$	8	$49\frac{3}{4}$
3	$43\frac{1}{8}$	9	$52\frac{7}{8}$
4	$45\frac{3}{4}$	10	$53\frac{5}{8}$
5	$47\frac{1}{2}$	11	$54\frac{1}{8}$
6	$49\frac{3}{8}$	12	$51\frac{5}{8}$

(a) Form a price index, using week 1 as the base.
(b) Form a price index, using week 7 as the base.

42. The following table shows the price and volume of shares (in thousands) traded for ABC Company and XYZ Company in the first 10 weeks of 1991.

	ABC Company		XYZ Company	
Week	Price	Volume	Price	Volume
1	$12\frac{2}{8}$	3.8	$25\frac{1}{8}$	6.4
2	$11\frac{3}{8}$	2.9	$24\frac{1}{4}$	7.1
3	$13\frac{5}{8}$	3.2	$24\frac{7}{8}$	6.9
4	$12\frac{7}{8}$	3.1	$26\frac{5}{8}$	7.4
5	$13\frac{1}{8}$	3.5	$26\frac{5}{8}$	7.7
6	$14\frac{2}{8}$	4.1	$25\frac{3}{8}$	6.9
7	$13\frac{7}{8}$	3.9	$27\frac{1}{8}$	7.4
8	$14\frac{1}{8}$	4.4	$27\frac{3}{8}$	6.8
9	$14\frac{7}{8}$	4.1	$28\frac{7}{8}$	7.2
10	$13\frac{7}{8}$	4.3	$29\frac{1}{8}$	7.4

Compute the market-value-weighted index using week 1 as the base.

43. Use the data given in question 42, and use week 1 as the base, to compute:

(a) The Laspeyres aggregative quantity index
(b) The Paasche aggregative quantity index
(c) Fisher's ideal quantity index

44. Redo question 43, this time using week 5 as the base. Compare the results to your results in question 43.

45. The following table shows the price per pound and the volume (in thousands of pounds) for chicken and beef at Eat More Grocery Store.

	Beef		Chicken	
Week	Price ($)	Volume	Price ($)	Volume
1	1.25	15	1.86	11
2	1.35	14	1.99	12
3	1.19	17	2.10	11
4	1.45	18	2.05	11
5	1.29	19	1.79	15
6	1.39	22	2.09	18
7	1.45	15	2.29	15
8	1.09	25	2.39	14
9	1.39	19	2.45	13
10	1.49	15	2.89	9

Use week 1 as the base to:

(a) Compute the Laspeyres aggregative price index.
(b) Compute the Paasche aggregative price index.
(c) Compute Fisher's ideal price index.

46. Use the data in question 45, and use week 1 as the base, to compute:

(a) The Laspeyres aggregative quantity index
(b) The Paasche aggregative quantity index
(c) Fisher's ideal quantity index

47. Use the week 10 data in question 46 to show that the ratio of the Laspeyres price index to the Laspeyres quantity index is equal to the ratio of the Paasche price index to the Paasche quantity index.
48. Show that the ratio of the Laspeyres price index to the Laspeyres quantity index is equal to the ratio of the Paasche price index to the Paasche quantity index.
49. Briefly explain the differences between aggregative price indexes and aggregative quantity indexes.
50. The following table shows the price and volume (in thousands) for shirts and pants at Snappy Dresser Department Store.

	Shirts		Pants	
Month	Price	Volume	Price	Volume
1	$27.00	50	$62.00	35
2	23.27	61	54.25	37
3	28.95	49	57.00	38
4	32.45	42	48.27	54
5	22.45	50	49.75	60
6	19.95	75	52.20	55
7	21.23	62	47.50	58
8	27.22	55	45.10	62
9	22.95	57	40.00	75
10	19.90	70	35.00	77
11	24.95	62	44.21	63
12	22.95	66	48.95	64

Compute the value indexes using week 12 as the base.

51. Use the data in question 50, and use week 1 as the base, to compute the value indexes.
52. Redo question 51, this time using month 6 as the base.
53. Using the business section of any newspaper, find the current value of the S&P 500 index. What stocks are included in the S&P 500? How is the index computed? Is the S&P 500 index a better or a worse measure of stock market activity than the Wilshire 5000?
54. Using the business section of any newspaper, find the most recent value of the Dow Jones Industrial Average (DJIA). What stocks are included in the DJIA?

Compare the DJIA to the Wilshire 5000 and the S&P 500 as a measure of stock market activity.

55. Explain the problems that would arise from comparing price indexes for computers over the last three decades.
56. The consumer price index (CPI) and the producer price index (PPI) are often reported as measures of the inflation rate. What problems appear using these indexes to measure the inflation rate? Do these two measures really indicate the "true" cost of living?
57. You are given the following information on prices and quantities for Knick-Knacks.

Year	Price	Quantity
1985	$21.00	12,300
1986	18.25	13,000
1987	19.31	21,300
1988	22.44	20,212
1989	23.51	24,345
1990	21.85	32,300

Using 1985 as the base year, compute the value price index for Knick-Knacks.

58. Using the data from question 57, compute the Paasche price index, again using 1985 as the base year.
59. Using your calculations from 57 to 58, compute the Laspeyres index. Why are you getting the same answer as in question 58?
60. For the data in question 57, are the Laspeyres quantity index and the Paasche quantity index equal? Why?
61. Suppose you are given the following information about wages and prices for 5 years.

Year	Annual salary	Consumer price index
1987	$38,202	95
1988	39,837	100
1989	41,001	108
1990	41,327	125
1991	55,943	200

(a) Compute the change in real salaries between 1988 and 1989.
(b) Are workers any better off in 1989 than they were in 1988?

62. Use the data given in question 61 to compute the change in real wages between 1987 and 1991. Are workers any better off?
63. Use the data presented in the following table to calculate the growth rate for nominal GNP, GNP deflator, and real GNP using MINITAB (*Hint*: Refer to Eqs. 19.17a, 19.17b, and 19.17c.)

Year	Nominal GNP	GNP deflators	Real GNP
1976	1,676.2	59.3	2,826.7
1977	1,991.1	67.3	2,958.6
1978	2,249.2	72.2	3.115.2
1979	2,509.2	78.6	3,192.4
1980	2,731.3	85.7	3,187.1
1981	3,053.9	94.0	3,248.8
1982	3,166.0	100.0	3,166.0
1983	3,407.0	103.9	3,279.1
1984	3,771.0	107.7	3,501.4
1985	4,013.1	110.9	3,618.7
1986	4,231.0	113.8	3,717.9
1987	4,514.4	117.4	3,845.3
1988	4,872.5	121.3	4,016.9
1989	5,200.7	126.3	4,117.7
1990	5,464.9	131.5	4,155.8

64. A quantity index measures a change in quantity from a base year to a particular year is the physical volume of production levels of industrial goods over time. Which of the following one is the most widely used and best-known quantity index in the United States?

    (a) Laspeyres quantity index
    (b) Producers index
    (c) FRB Index of Industrial Production
    (d) Paasche quantity index

65. The Laspeyres index is a weighted aggregate price index where the weight for each item is its

    (a) Base-year price
    (b) Base-year quantity
    (c) Current-year price
    (d) Current-year quantity

66. The Paasche index is a weighted aggregate price index where the weight for each item is its

    (a) Base-year price
    (b) Base-year quantity
    (c) Current-year price
    (d) Current-year quantity

67. What is the disadvantage of the CPI?

    (a) It tends to give less weight to those items that show a dramatic price increase.
    (b) It tends to give more weight to those items that show a dramatic price increase.

(c) It tends to give more weight to those items that show a dramatic price decrease.
(d) It tends to give the same weight to those items that show a dramatic price decrease.

68. Use the data in the following table to calculate GNP deflator, the growth rate for nominal GNP, GNP deflator, and real GNP.

Year/quarter	Real GNP ($)	Nominal GNP ($)
2005 Q1	10,941.90	12,468.70
2005 Q2	11,014.70	12,596.20
2005 Q3	11,151.20	12,835.70
2005 Q4	11,151.10	12,979.80
2006 Q1	11,294.00	13,242.00
2006 Q2	11,362.50	13,406.00
2006 Q3	11,375.90	13,494.70
2006 Q4	11,447.80	13,655.90
2007 Q1	11,466.70	13,828.80
2007 Q2	11,580.00	14,056.30
2007 Q3	11,744.60	14,270.90
2007 Q4	11,799.10	14,451.80

Appendix 1: Options on Stock Indices and Currencies[6]

Index Options

Index option is the option on stock index instead of individual stocks as discussed in Appendix 2 of Chap. 6, Appendices 2 and 3 of Chap. 7, and Appendix 4 of Chap. 13. Many different index options currently trade in the United States. The most popular contracts are those on the S&P 500 Index and the S&P 100 Index (CBOE).

Index options may be European or American. For example, the contract on the S&P 500 is European, whereas that on the S&P 100 is American. One CBOE contract is to buy or sell 100 times the index at the specified strike price of 280. If it is exercised when the value of the index is 292, the writer of the contract pays the holder $(292 - 280) \times 100 = \$1{,}200$. This cash payment is based on the index value at the end of the day in which the exercise instructions are issued. Not surprisingly, investors usually wait until the end of a day before issuing these instructions.

[6] See Hall, J.C.: *Introduction to Futures and Options Markets*, 3rd edn. Prentice-Hall, New Jersey (1998), Chap. 12 for further discussion on this topic.

In valuing index, assume that it could be treated as a security paying a known dividend yield. Therefore, the European style of index call options can be evaluated in terms of the European style of stock call option formula defined as

$$C = S'N(d_1) - Xe^{-rT}N(d_2) \tag{19.18}$$

where $S' = Se^{-qT}$, q = dividend yield and S = value of index.

$$d_1 = \frac{\left[l_n(S/X) + \left(r - q + \frac{1}{2}\sigma^2\right)T\right]}{\sigma\sqrt{T}}$$

$$d_2 = d_1 - \sigma\sqrt{T}$$

See Eq. 7.35 in Appendix 2 of Chap. 7 for the definitions of all other variables.

Example 19.1 Index Option Valuation. Consider a European call option on the S&P 500 that is 2 months from maturity. The current value of the index is 950, the exercise price is 900, the risk-free interest rate is 6 % per annum, and the volatility of the index is 15 per annum. Dividend yields of 0.2 % and 0.3 % are expected in the first month and the second month, respectively. In this case $S = 950$, $X = 900$, $r = 0.06$, $\sigma = 0.15$, and $T = 2/12$. The total dividend yield during the option's life is 0.2 + 0.3 = 0.5 %. This is 3 % per annum. Hence, $q = 0.03$ and

$$d_1 = \frac{\ln(950/900) + \left[0.06 - 0.03 + (0.15)^2(0.5)\right]\left(\frac{2}{12}\right)}{(0.15)\sqrt{\frac{2}{12}}}$$

$$= \frac{0.054 + 0.007}{0.061} = 1$$

$$d_2 = 1 - (0.15)\sqrt{\frac{2}{12}} = 0.93$$

From Table A.3, we obtained that

$$N(d_1) = 0.8413 \quad N(d_2) = 0.8238$$

so that the call price, C, is given by Eq. 19.18

$$C = 950(0.8413)e^{-0.03\times 2/12} - 900(.8238)e^{-0.006\times 2/12} = 795.24 - 734.01$$
$$= 61.23$$

One contract would cost \$61.23 × 100 = \$6123.

If the absolute amount of the dividend that will be paid on the stocks underlying the index (rather than the dividend yield) is assumed to be known, the basic

Black–Scholes formula can be used with the initial stock price reduced by the present value of the dividends. This is the approach recommended in Appendix 4 of Chap. 13 for a stock paying known dividend.

Currency Option

Currency option is option on spot exchange rate instead of either individual stock or stock index. An exchange rate is the price of one currency in terms of another currency. For example the exchange rate between the Japanese yen and US dollar is 130.77 on December 15, 1997.

The valuation model for the European type of currency call option can be defined as

$$C = S^{-r_f T} N(d_1) - X e^{-rT} N(d_2) \tag{19.19}$$

where
S = spot exchange rare, r = domestic risk-free rate, T= term to maturity in years
r_f = foreign risk-free rate, X = exercise price
σ = standard deviation of spot exchange rate

$$d_1 = \frac{\left[\ln\left(\frac{S}{X}\right) + \left(r - r_f + \frac{\sigma^2}{2}\right)T\right]}{\sigma\sqrt{T}}$$

$$d_2 = d_1 - \sigma\sqrt{T}$$

By comparing Eq. 19.19 with Eq. 19.18, it is found that Eq. 19.19 can be obtained by replacing q by r_f.

Example 19.2 Valuation of Currency Option. Consider a 4-month European call option on the Japanese yen. Suppose that the current exchange rate is 130, the exercise price is 125, the risk-free rate in the United States is 6 % per annum, and the risk-free rate in Japan is 2 % per annum. The volatility of foreign exchange rate is 15 %.

Substituting $S = 130$, $X = 125$, $r = 0.06$, $r_f = 0.02$, $\sigma = .15$, and $T = 4/12$ into Eq. 19.19, we obtain

$$d_1 = \frac{\ln\left(\frac{130}{125}\right) + \left[0.06 - 0.02 + \left(\frac{(0.15)^2}{2}\right)\right]\left(\frac{4}{12}\right)}{(0.15)\sqrt{\frac{4}{12}}} = \frac{0.0392 + 0.0171}{0.0866} = 0.6501$$

$$d_2 = 0.6501 - (0.15)\sqrt{\frac{4}{12}} = 0.5635$$

From Table A.3, we obtain

$$N(0.65) = 0.7422, \quad N(0.56) = 0.7123$$

Substituting all related information into Eq. 19.19, we obtain

$$\begin{aligned} C &= (130)e^{-\frac{0.02}{3}}(0.7422) - (125)e^{-\frac{0.06}{3}}(0.7123) \\ &= 95.8395 - 87.2746 \\ &= 8.5649 \end{aligned}$$

Appendix 2: Index Futures and Hedge Ratio

The most exciting financial innovation of the 1980s might be the introduction of stock index futures contracts. These contracts, written on the value of various stock index portfolios as defined in the text of this chapter, provide important benefits to stock portfolio managers.

The first stock index futures contract was introduced in February 1982 by the Kansas City Board of Trade. This Value Line futures contract is written on the Value Line Composite Index, a stock index that consists of approximately 1,700 stocks from the New York, American, and OTC stock markets. The Chicago Mercantile Exchange quickly followed suit in April 1982 with a futures contract on the S&P 500 stock index, and then the Chicago Board of Trade in July 1984 followed with a futures contract on the Major Market Index.

Most recently, the Chicago Board of Trade introduce Dow Jones Index Futures in October 1997.

Stock portfolio manager can use index futures to hedge his (or her) portfolio. Now, we discuss how the minimum variance type of hedge ratio can be derived in accordance with method used to derive the optimal weights of a portfolio which has been discussed in Appendix 1 of Chap. 13. We first define

ΔS: Change in spot price, S, during a period of time equal to the life of the hedge
ΔF: Change in futures price, F, during a period of time equal to the life of the hedge
σ_S: Standard deviation of ΔS
σ_F: Standard deviation of ΔF
ρ: Coefficient of correlation between ΔS and ΔF
h: Hedge ratio

When the hedger is long for the asset and short futures, the change in the value of the hedger's position during the life of the hedge is

$$\Delta S - h\Delta F$$

For a long hedge, it is

$$h\Delta F - \Delta S$$

In either case the variance, σ, of the change in value of the hedged position is given by

$$\sigma = \sigma_S^2 + h^2\sigma_F^2 - 2h\rho\sigma_S\sigma_F$$

so that

$$\frac{\partial\sigma}{\partial h} = 2h\sigma_F^2 - 2\rho\sigma_S\sigma_F$$

Setting this equal to zero and noting that $\partial^2\sigma/\partial h^2$ is positive, we see that the value of h which minimizes the variance is

$$h = \rho\frac{\sigma_S}{\sigma_F} \tag{19.20}$$

The optimal hedge ratio is therefore the product of the coefficient of correlation between ΔS and ΔF and the ratio of the standard deviation of ΔS to the standard deviation of DF.

If $\rho = 1$ and $\sigma_F = \sigma_S$, the optimal hedge ratio, h, is 1.0. This is to be expected since in this case the futures price mirrors the spot price perfectly. If $\rho = 1$ and $\sigma_F = 2\sigma_S$, h is 0.5. This result is also expected since in this case the futures price always changes by twice as much as the spot price.

The optimal hedge ratio, h, defined in Eq. 19.20 can be estimated by using the following regression

$$\Delta S_t = a_0 + a_1\Delta F_t + e_t \tag{19.21}$$

where ΔS_t = change in spot price in period t

ΔF_t = change in futures price in period t

a_0 and a_1 are the intercept and slope of a regression, respectively. The estimated a_1 is the estimated hedge ratio, h. Application of Eq. 19.21 has been shown in Problem 60 of Chap. 14.

Chapter 20
Sampling Surveys: Methods and Applications

Chapter Outline

Key Terms

Sampling error	Simple random sampling
Nonsampling errors	Random number table
Sample selection bias	Finite sample adjustment factor
Response bias	Stratified random sampling
Measurement error	Optimal allocation of sample
Nonresponses	Two-stage cluster sampling
Self-selection bias	Jackknife method

20.1 Introduction

In statistics, we are interested in information about a population. For example, we might be interested in how the residents of a community feel about the construction of a new high school. There are two ways to obtain information about how the residents feel about this issue. We could take a census and simply ask each and

C.-F. Lee et al., *Statistics for Business and Financial Economics*, 1019
DOI 10.1007/978-1-4614-5897-5_20, © Springer Science+Business Media New York 2013

every resident about his or her attitude toward such a project. Or we could take a smaller sample of the residents and try to draw inferences about the community's feelings from the feelings that members of this sample express.

In Part III of this book, we investigated sampling in terms of only simple random sampling, in which each potential sample of N members has an equal chance of being chosen. However, survey sampling is more likely to require elaborate sampling designs for the selection of sample members; these designs are discussed in Sects. 20.3 and 20.5. We also focus in this chapter on the problem confronting researchers who want to discover something about a population that is not very large. (In previous chapters, we generally assumed that the number of population members was very large compared with the number of sample members.) And we discuss the advantages and disadvantages of sampling and show how sampling is applied to decision making in business and economics.

As we noted earlier, there are four reasons why sampling may be preferred to taking a census. First, sampling is more economical. Second, it is preferable when information needs to be gathered quickly. A third reason for using a sample instead of a census is that the population of interest may be very large. A fourth reason is quality control, which we discussed in detail in Sects. 10.8 and 10.9 of Chap. 10.

This chapter discusses techniques for designing a sampling experiment, and it adds examples and applications to our earlier (Chap. 1) discussion of sampling. First, the basic sampling methods of simple random sampling and stratified sampling are explained. Then we address determining the sample size and discuss sampling and nonsampling errors. Two-stage cluster sampling is also investigated, and we compare ratio estimation and regression estimation. Finally, applications of sampling methods are demonstrated. Appendix 1 shows how the jackknife method is used to remove the bias of a sample estimator.

20.2 Sampling and Nonsampling Errors

There are several advantages of sampling over taking a census. However, working with a sample taken from a population does not enable us to determine the precise value of the population parameter, such as population mean or variance. This kind of error is due to sampling error. *Sampling error* is the difference between the sample estimate and its population parameter that is due entirely to the fact that sample instead of census data are used to estimate the parameter. For instance, say we were interested in determining the mean income of lawyers in a particular law firm. Had we used a simple random sample consisting of 50 lawyers, the difference between the mean income of this sample and population mean income would be the sampling error.[1] If by chance our sample consisted entirely of partners in the law firm, our resulting error would be quite large. Now although it is unlikely that a

[1] We will return to this example in Example 20.3 in the next section of this chapter.

random sample of 50 lawyers from a law firm consisting of 475 associates and 50 partners would result in the selection of only partners, it is theoretically possible. And if it happened, our inferences about the mean income of this law firm would be based on the income of the partners alone – and hence would be greatly overstated. Parts of the errors in estimating the population parameters that result from "the luck of the draw" are not sampling errors anymore.

Errors that are unconnected with the pure random sampling procedure used fall into the category of *nonsampling errors*. Selection of the wrong population is one example of nonsampling error. For example, a researcher who tries to draw inferences about the views of Americans on gun control would fall victim to nonsampling error if he chose to sample only members of the National Rifle Association. This kind of nonsampling error is called *sample selection bias*.

Another source of nonsampling error is *response bias*. Poorly worded questionnaires and improper interviewing techniques may distort the responses of individuals so much that they do not accurately reflect the respondents' true opinions. Furthermore, respondents may have an incentive to distort the truth – say, to exaggerate their incomes if they think their friends will have access to the survey or to understand their incomes if they think the IRS will find out. This kind of error is called *measurement error*.

A third possible source of nonsampling error is *nonresponses*. Individuals who choose not to respond to a survey may have very different views from those who do respond. For example, automobile owners who are dissatisfied with their cars may be more likely than satisfied customers to respond to a questionnaire on customer satisfaction. This kind of bias is called *self-selection bias*.

Because nonsampling errors can have a great impact on the results of a survey, the researcher must design the study carefully to minimize these errors. In a study on the obtaining of periodic market information on small business for four large cities, Keon and Assael (1982) found that most of the time, nonsampling errors count for more than 90 % of the total errors![2]

20.3 Simple and Stratified Random Sampling

20.3.1 Designing the Sampling Study

The first step in sampling is to design a study that will yield the information the researcher needs. Designing the study involves determining what questions to address and identifying the population that will make possible achievement of the study's goals. Poor planning may add to the costs of the study or even invalidate the results.

[2] J. K.eon and H. Assael (982), "Nonsampling vs. Sampling Errors in Survey Research" *Journal of Marketing*, Spring 1982, 114–123.

There are six main steps in survey sampling:

1. Determine what information is required for the study.
2. Construct a population list to be used for the sampling survey.
3. Decide what method to use in selecting the sample.
4. Determine the sample size.
5. Decide by what method to infer population parameters from sample data.
6. Draw appropriate conclusions from the sample information.

We will discuss the last four of these steps.

20.3.2 Statistical Inferences in Terms of Simple Random Sampling

Once the researcher has determined the questions to be addressed and the population to be studied, the data collection process begins. The issue now is how the sample members should be selected from the population.

The easiest sampling technique is *simple random sampling,* in which each member of a population has an equal and independent chance of being chosen. For example, suppose a population consists of 100 balls, numbered from 1 to 100, that represent households to be surveyed for their annual income. If we were interested in a random sample of 10 balls from this population, we could simply place all 100 balls in a bag, mix the balls thoroughly, and draw 10 of them.

Alternatively, as noted in Sect. 8.2, we could use a table of random numbers to achieve the same objective more efficiently.

20.3.2.1 Random Number Tables

Random numbers were discussed in detail in Chap. 8. We now consider *random number tables*. There are several random number tables, but we chose to reproduce, as Table A8 in Appendix A, part of the Rand Corporation table, which has one million digits.

Let us illustrate the use of random number tables with an example. Suppose there are 800 students at the Rutgers University School of Business, and we wish to select a random sample of 15 students to estimate their average grade. A list of students is compiled, and each is assigned a serial number from 001 to 800.

Because 800 is a 3-digit number, we should list only 3 digits of the random numbers. The procedure is started at some arbitrary point on Table A8 – say, the top of the third column – and the last 3 digits of the random numbers are read off. Thus, the first 20 are

769	463	779	850
630	179	596	562
240	238	742	384
610	061	976	951
127	201	033	221

Hence, we select the following 15 numbers for our sample:

769	463	779
630	179	596
240	238	742
610	061	033
127	201	562

As you have noticed, only numbers less than 800 were selected. If we needed to choose 50 students, we would continuously choose numbers smaller than 800 until 50 students had been selected. During the selection procedure, we didn't replace any number that had already been chosen. This is known as sampling without replacement. Sampling with replacement, which allows the possibility of an individual being included in the sample more than once, will not be discussed here.

After selecting the random numbers from the table, we could make them *all* usable by subtracting 800 from those greater than 800. Hence, for the random number 976, the 976–800 = 176th student is selected.

20.3.2.2 Confidence Interval for Population Mean

We will use mean and standard deviation statistics and the central limit theorem (introduced in Chap. 8) to draw inferences about the total population on the basis of information gathered from a random sample. To use the central limit theorem, we must assume that the sample is sufficiently large. However, when the population size N is finite, we use an adjustment factor to obtain the unbiased estimator.

Let $x_1, x_2, \ldots, x_n$ denote the values observed from a simple random sample of size n taken from a population of N numbers with mean μ. Then estimating the population mean involves the following steps:

1. Calculate the sample mean $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{20.1}$$

$\bar{x}$ is an unbiased estimator of the population mean μ if $E(x_i) = \mu$ for all i.

2. Calculate the variance of $\bar{x}$:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{20.2}$$

The sample variance s^2 is a biased estimator of the population variance σ^2 when the sample size is finite. Hence, the unbiased estimated variance for the sample mean is[3]

$$\hat{\sigma}_{\bar{x}}^2 = \frac{s^2}{n} \cdot \frac{(N-n)}{N-1} \tag{20.3}$$

where N and n are population size and sample size, respectively. $\frac{N-n}{N-1}$ is called the *finite sample adjustment factor*. $\sigma_{\bar{x}}$ is the standard deviation of $\bar{x}$. If N is large, then the adjustment factor approximately equals 1.

3. Because we have taken a sample of n from the population consisting of N members, we cannot be certain of the true population mean. However, if the sample is large enough to permit use of the central limit theorem, we can construct a $100(1-\alpha)$ percent confidence interval for the population mean. The confidence interval will be

$$\bar{x} - Z_{\alpha/2}\hat{\sigma}_{\bar{x}} < \mu < \bar{x} + Z_{\alpha/2}\hat{\sigma}_{\bar{x}} \tag{20.4}$$

where $Z_{\alpha/2}$ is the number for which

$$P\left[Z > Z_{\alpha/2}\right] = \alpha/2$$

and the random variable Z follows a standard normal distribution.

In other words, Eq. 20.4 enables us to construct an interval estimate for the true mean of the population, as illustrated in Fig. 20.1.

Example 20.1 Simple Random Sampling to Determine Household Income. Suppose an investment adviser is trying to decide whether a small retirement community consisting of 1,000 residents represents a promising source of potential clients. To determine the potential business, the investment adviser decides to analyze the size of the residents' investment portfolios. A random sample of 75 residents, who were able to respond anonymously, produces a sample mean of $375,000 with a sample standard deviation of $120,000.

We can use this information to construct a 95 % confidence interval for the mean value of the investment portfolio:

[3] This result is obtained under the assumption that $\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2/N$. If $\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2/N - 1$, then the finite sample adjustment factor will be $(N-n)/N$. This kind of finite sampling adjustment factor will be used in both stratified random sampling and two-staged cluster sampling.

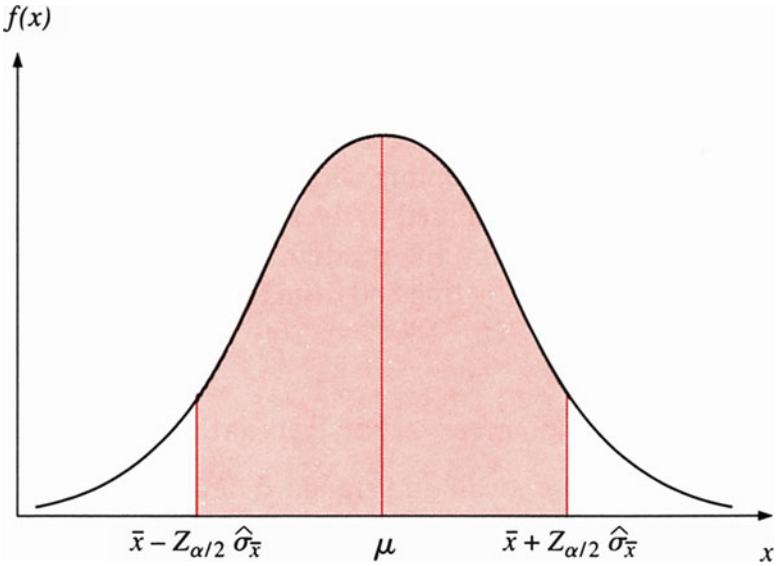


Fig. 20.1 Confidence interval estimate for population mean

$$
\begin{aligned}
N &= 1{,}000 \\
n &= 75 \\
\bar{x} &= \$375{,}000 \\
s &= \$120{,}000
\end{aligned}
$$

First, we need to produce an unbiased estimate of the standard deviation of the sample mean from Eq. 20.3:

$$
\begin{aligned}
\hat{\sigma}^2_{\bar{x}} &= \frac{s^2}{n} \cdot \frac{N-n}{N-1} \\
&= \frac{(120{,}000)^2}{75} \cdot \frac{1000-75}{1000} \\
&= 177{,}772{,}800 \\
\hat{\sigma}_{\bar{x}} &= \sqrt{177{,}772{,}800} = 13{,}333.15
\end{aligned}
$$

A 95 % confidence interval can be constructed as follows:

$$\bar{x} - Z_{.025}\hat{\sigma}_{\bar{x}} < \mu < \bar{x} + Z_{.025}\hat{\sigma}_{\bar{x}}$$

The value for $Z_{.025}$ can be found in Table A3 of Appendix A. It is $Z_{o:,} = 1.96$, so the 95 % confidence interval for the mean value of the investment portfolio for this population is

$$375{,}000 - (1.96)(13{,}333.15) < \mu < 375{,}000 + (1.96)(\$13{,}333.15)$$

or

$$\$348{,}867.03 < \mu < \$401{,}132.97$$

Given the information from the sample, we may expect, with 95 % confidence, that the true population mean μ falls between \$348,867.03 and \$401,132.97.

20.3.2.3 Confidence Interval for Population Proportion

The same approach we used to calculate a confidence interval for a random sample with mean x can be applied to sample proportions. Again, we follow these three steps:

1. Compute the sample proportion $\hat{p}$, which is an unbiased estimator of the population proportion p.
2. Compute an unbiased estimator for the variance of the estimator:

$$\hat{\sigma}_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1} \tag{20.5}$$

3. If the sample is large enough, use the central limit theorem to construct a $100(1 - \alpha)$ percent confidence interval:

$$\hat{p} - Z_{\alpha/2}\hat{\sigma}_{\hat{p}} < \hat{p} + Z_{\alpha/2}\hat{\sigma}_{\hat{p}} \tag{20.6}$$

Equation 20.6 can be interpreted to mean we may expect, with $100(1 - \alpha)$ percent confidence, that the population proportion p falls within this interval.

Example 20.2 Simple Random Sampling to Determine the Proportion of College-Bound High School Seniors. Suppose we want to determine the proportion of college-bound high school seniors in a class of 500. A survey of 30 randomly selected students reveals that 19 will be attending college. Given this information, we want to estimate the population proportion p:

$$N = 500$$
$$n = 30$$
$$\hat{P} = \frac{19}{30} = .6333$$

To estimate our confidence interval, we need to calculate an unbiased estimate of the variance of the population proportion:

$$\begin{aligned}\hat{\sigma}_{\hat{p}}^2 &= \frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1} \\ &= \frac{.6333(1-.6333)}{30} \cdot \frac{500-30}{500-1} \\ &= .00729 \\ \hat{\sigma}_{\hat{p}} &= \sqrt{.00729} = .0854\end{aligned}$$

The $100(1-\alpha)$ percent confidence interval in terms of Eq. 20.6 is

$$\hat{p} - Z_{\alpha/2}\hat{\sigma}_{\hat{p}} < p < \hat{p} + Z_{\alpha/2}\hat{\sigma}_{\hat{p}}$$

To construct a 90 % confidence interval, we use the $Z_{.05}$-value given in Table A3 of Appendix A. It is $Z_{.05} = 1.645$, so the 90 % confidence interval can be given as

$$\begin{aligned}.6333 - (1.645)(.0854) < p < .6333 + (1.645)(.0854) \\ .4928 < p < .7738\end{aligned}$$

This implies that we may expect, with 90 % confidence, that the true proportion of high school seniors who will be attending college falls between 49.28 % and 77.38 %.

20.3.3 Stratified Random Sampling

There are times when simple random sampling is not the best sampling method. In some cases, it may be more appropriate to divide the population into groups or *strata*. *Stratified random sampling* is the selection of independent simple random samples from each stratum of the population.

A stratified random sample may be preferable to a simple random sample when there is reason to believe that different groups within the population have markedly different views. For example, suppose a researcher at Johnson & Johnson is interested in employees' opinions on a child care program. He believes that the views of female employees are important and that their views may be quite different from those of male employees. If the company has a high percentage of male employees, a simple random sample may not guarantee that the sample percentage of female employees is the same as the population percentage. In this instance, a stratified random sampling is called for, wherein the population is first divided into male and female subgroups, or strata, and a simple random sample is taken from each stratum.

To conduct a stratified random sampling survey, we begin by dividing the total population of N members into H mutually exclusive and collectively exhaustive groups. Each of these H strata contains its own population consisting of $N_1, N_2, \ldots, N_H$ members. Because the strata are mutually exclusive and collectively exhaustive, we know that

$$N_1 + N_2 + \ldots + N_H = N$$

Our approach in stratified random sampling is to treat each stratum as a separate population, and our sampling survey consists of sampling each stratum separately. Using this technique, we will take a sample of $n_1, n_2, \ldots, n_H$ from each stratum. The samples taken from the strata do not have to be the same size. The total sample taken is the sum of all these samples; that is,

$$n = n_1 + n_2 + \ldots + n_H$$

The techniques for stratified random sampling can be used to

1. Produce an unbiased estimator for the population mean μ.
2. Produce an unbiased estimator for the variance of the sample mean.
3. Construct a $100(1 - \alpha)$ percent confidence interval for the population mean.

To produce an unbiased estimator for the population mean, we take a weighted average of the sample means in the individual strata:

$$\bar{x}_{st} = \sum_{j=1}^{H} W_j \bar{x}_j \tag{20.7}$$

where

$\bar{x}_{st}$ = sample mean for the overall population from stratified sampling

$W_j = \frac{N_j}{N}$ = proportion of the jth stratum

$\bar{x}_j$ = sample mean for the jth stratum

Next we need to find an unbiased estimator of the variance of the sample mean. An unbiased estimator of the variance of the sample mean for the jth stratum is

$$\hat{\sigma}^2_{\bar{x}_j} = \frac{s_j^2}{n_j} \cdot \frac{N_j - n_j}{N_j}$$

where s_j^2 is the sample variance for the jth stratum. Note that this formula is identical to the estimator we used in simple random sampling. Because each stratum is treated as a separate population, the variance for each is calculated in the same way as in simple random sampling.

To find an unbiased estimator of the variance of the estimator μ, we again take a weighted average of the variances of the individual strata[4]:

[4] $\bar{x}_{st} = \sum_{J=1}^{H} W_j \bar{x}_j$, where $W_j = N_j/N_j$. Because the samples N_j are selected by random sampling and are independent of each other, Eq. 20.8 holds.

$$\hat{\sigma}^2_{\bar{x}_{st}} = \sum_{j=1}^{H} W_j^2 \hat{\sigma}^2_{\bar{x}_j} \tag{20.8}$$

To construct confidence intervals for the population mean, we again need a sample size large enough for us to assume normality. The confidence interval is

$$\bar{x} - Z_{\alpha/2}\hat{\sigma}_{\bar{x}_{st}} < \mu < \bar{x} + Z_{\alpha/2}\hat{\sigma}_{\bar{x}_{st}}$$

Example 20.3 Stratified Random Sampling to Determine the Mean Income of Lawyers. Suppose a researcher is interested in determining the mean income of lawyers at a large New York City law firm. The firm consists of 525 lawyers, 475 associates, and 50 partners. Because there are relatively few partners in the population, the researcher believes a simple random sample might understate the earnings of the partners. He decides to undertake a stratified random sample of 75 lawyers: 50 associates and 25 partners. The sample means and standard deviations are

Associates	Partners
$\bar{x}_1 = \$62{,}750$	$\bar{x}_2 = \$271{,}860$
$S_1 = \$11{,}620$	$S_1 = \$80{,}210$
$n_1 = 50$	$n_2 = 25$
$N_1 = 475$	$N_2 = 50$

An unbiased estimator for the population mean income can be calculated as

$$\bar{x} = \sum_{j=1}^{2} W_j \bar{x}_j$$

where

$$\begin{aligned} W_j &= \frac{N_j}{N} \\ \bar{x} &= \frac{475}{525}(62{,}750) + \frac{50}{525}(271{,}860) \\ &= \$82,665 \end{aligned}$$

To produce an unbiased estimator of the variance of the estimator for μ, we first need to produce unbiased variance estimators for all the strata.

$$\begin{aligned}
\sigma^2_{\bar{x}_1} &= \frac{(11{,}620)^2}{50} \cdot \frac{475-50}{475} \\
&= 2{,}416{,}226 \\
\sigma^2_{\bar{x}_2} &= \frac{(80{,}210)^2}{25} \cdot \frac{50-25}{50} \\
&= 128{,}672{,}882 \\
\hat{\sigma}^2_{\bar{x}} &= \sum_{j=1}^{2} W_j^2 \hat{\sigma}_{\bar{x}_j}{}^2 \\
&= \left(\frac{475}{525}\right)^2 (2{,}416{,}226) + \left(\frac{50}{525}\right)^2 (128{,}672{,}882) \\
&= 3{,}145{,}009 \\
\hat{\sigma}_{\bar{x}} &= \sqrt{3.145.009} = 1.773
\end{aligned}$$

For a 95 % confidence interval, we use the $Z_{.025}$-value from Table A3 in Appendix A. It is $Z_{.025} = 1.96$, so the 95 % confidence interval for the income of lawyers at this law firm is

$$\begin{aligned}
82{,}665 - (1.96)(1{,}773) < \mu < 82{,}665 + (1.96)(1.773) \\
\$79{,}190 < \mu < \$86{,}140
\end{aligned}$$

We can say with 95 % confidence that the population mean μ falls between \$79,190 and \$86,140.

20.4 Determining the Sample Size

We have discussed the advantages of sampling overtaking a census, and we have looked at two important sampling methods. One fundamental question remains unanswered: How large should the sample be?

20.4.1 Sample Size for Simple Random Sampling

For simple random sampling, the sample size n can be found via Eq. 20.3, the formula for finding the variance of the estimator $\bar{x}$:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{(N-1)}$$

By solving for n, we determine the sample size. We then multiply both sides of the equation by $(N-1)n$:

$$(N-1)n\sigma_{\bar{x}}^2 = N\sigma^2 - n\sigma^2$$

Next we add $n\sigma^2$ to both sides of the equation and rearrange:

$$\left[(N-1)\sigma_{\bar{x}}^2 + \sigma^2\right]n = N\sigma^2$$

Dividing both sides by the bracketed term yields

$$n = \frac{N\sigma^2}{(N-1)\sigma_{\bar{X}}^2 + \sigma^2} \tag{20.9}$$

If the population variance σ^2 is known, Eq. 20.9 can be used to determine the sample size necessary to achieve any specified value for the level of precision $\bar{x}$, $\sigma_{\bar{x}}^2$. Eq. 20.9 makes apparent the inverse relationship between $\sigma_{\bar{x}}^2$ and n; that is, the smaller the variance of the estimator that we desire, the larger our sample needs to be.

Example 20.4 Sample Size for Accounts Receivable, Case I. Crow Company's accountant decides that the best way to determine the company's mean accounts receivable is to take a simple random sample of the 1,025 accounts. Assume that the population variance σ^2 is \$2,425. What size sample should the accountant take if she would like to have a level of precision, as measured by $\sigma_{\bar{x}}^2$, of \$75?

$$\begin{aligned} n &= \frac{N\sigma^2}{(N-1)\sigma_{\bar{x}}^2 + \sigma^2} \\ &= \frac{1025(2{,}425)}{(1025-1)(75) + 2{,}425} = 31.37 \end{aligned}$$

That is, a simple random sample of 32 accounts receivable will produce the desired result. Note that we rounded the sample size up to the nearest whole number.

Example 20.5 Sample Size for Accounts Receivable, Case 2. Use the information from Example 20.4 but assume the accountant would like $\sigma_{\bar{x}}^2 = \$50$:

$$\frac{1{,}025(2{,}425)}{(1{,}025-1)(50) + 2{,}425} = 46.35$$

The accountant must increase the sample size to 47 if she wishes to reduce $\sigma_{\bar{x}}^2$ to \$50 or—what is the same thing—to improve her precision from \$75 to \$50.

Example 20.6 Sample Size for Accounts Receivable, Case 3. Suppose our accountant is not sure what value for $\sigma_{\bar{x}}^2$ would be appropriate. However, she would like to produce a sample in which the 95 % confidence interval extends \$10 on each side of

the sample mean. She can do this by noting that $Z_{\alpha/2} \cdot \sigma_{\bar{x}} =$ length of the confidence interval on each side of the sample mean. Thus, for a 95 % confidence interval extending \$10 on each side of the mean, $1.96\ \sigma_{\bar{x}} = \$10$. Solving for $\sigma_{\bar{x}}$, we get

$$\sigma_{\bar{x}} = \frac{10}{1.96} = 5.10$$
$$\sigma_{\bar{x}}^2 = (5.10)^2 = 26.01$$

so

$$n = \frac{1025(2{,}425)}{(1025-1)(26.01) + 2{,}425}$$
$$= 85.54$$

In this instance, the accountant randomly samples 86 accounts receivable.

If we define the absolute value of the difference between sample mean $\bar{x}$ and population mean μ ($d = |\bar{x} - \mu|$) as a precision measure, then it can be shown that the relationship among precision, the level of reliability (the Z-value of a normal distribution), and standard error $\sigma_{\bar{x}}$ is $d = z\ \sigma_{\bar{x}}$. From this relationship, it can be shown that the sample size can be defined as

$$n = \frac{(z\sigma)^2}{d^2} \tag{20.10a}$$

$$n = \frac{N(z\sigma)^2}{(N-1)d^2 + (z\sigma)^2} \tag{20.10b}$$

where Eqs. 20.10a and 20.10b are for sampling with replacement and without replacement, respectively.

Using Eq. 20.10b, we can calculate the sample size for Example 20.6 as

$$n = \frac{(1025)(1.96)^2(2425)}{(1025-1)(10)^2 + (1.96)^2(2425)} = 85.47$$

The sample size calculated from Eq. 20.10b is almost identical to that of Example 20.6.

Now let's consider simple random sampling for estimating the population proportion p. Let $\hat{p}$ be the random variable that represents the sample proportion. Then, from Eq. 20.5, we can solve this equation for sample size. We obtain

$$n = \frac{N\hat{p}(1-\hat{p})}{(N-1)\sigma_{\hat{p}}^2 + \hat{p}(1-\hat{p})} \tag{20.11}$$

The sample size obtained from Eq. 20.11 does not connect with the desired degree of precision. To directly relate the desired precision with the estimate of sample size, we first let d represent a "margin of error" in estimating sample proportion $\hat{p}$. Then we also let α represent the risk that actual error is larger than d. In other words,

$$P(|\hat{p} - p| \geq d) = \alpha$$

The formula for a sample size that uses the information of d is derived as follows: First, we let

$$d^2 = z^2 \frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1}$$

where z is the abscissa of the normal curve that cuts off an area α at the tails. Solving for n yields

$$n = \frac{Nz^2\hat{p}(1-\hat{p})}{(N-1)d^2 + z^2\hat{p}(1-\hat{p})} \tag{20.12'}$$

Both Eqs. 20.11 and 20.12 involve the unknown population proportion p whose estimation is the objective of the study. A conservative estimate of n is obtained by choosing for p the value nearest to $\frac{1}{2}$ in the range in which p is thought likely to lie.

Example 20.7 Nielson Survey About the Evening News on NBC. Suppose the Nielson organization is planning to make a simple random sampling to estimate what percentage of American TV viewers watch the 6:30 evening news on NBC. What is the sample size needed for $d = .04$?

The potential number of TV watchers, N, is very large, so Eq. 20.11 can be approximated by

$$n' = \frac{[z^2\hat{p}(1-\hat{p})]/d^2}{1 + 1/N\left(\frac{z^2\hat{p}(1-\hat{p})}{d^2} - 1\right)} = \frac{z^2\hat{p}(1-\hat{p})}{d^2} \tag{20.12'}$$

Substituting $d = .04$, $\hat{p} = .5$, and z = 1.96 into Eq. 20.12′, we obtain

$$n' = \frac{(1.96)(.5)(.5)}{(.04)^2} = 600.25$$

A simple random sample of 601 will suffice.

20.4.2 Sample Size for Stratified Random Sampling

We can also derive a formula for the sample size needed in stratified random sampling. As in the case of simple random sampling, our required sample size depends on the variance of the population and the desired level of precision. However, the size of the sample in stratified random sampling also depends on one other factor: how we allocate the total sample among the strata. There are two possible approaches:

1. *Proportional allocation*. In cases where the sampling will be distributed proportionally, the proportions are determined by the relative sizes of the strata. For example, if the total population is 1,000 and the total population of the first stratum is 400, 40 % of the total sample is assigned to the first stratum:

$$n_j = \frac{N_j}{N} n$$

For a proportional allocation from a stratified sample, the sample size n can be determined by the formula[5]

$$n = \frac{\sum_{j=1}^{H} N_j s_j^2}{N\sigma_{\bar{x}_{st}}^2 + \frac{1}{N}\sum_{j=1}^{H} N_j s_j^2} \tag{20.13}$$

where $\sigma_{\bar{x}_{st}}^2$ is the desired variance and s_j^2 is the sample variance in the jth stratum. From Eq. 20.13, we can see that there is an inverse relationship between the sample size n and the degree of precision we desire, as measured by $\sigma_{\bar{x}}^2$.

2. *Optimal allocation with similar variable cost in each stratum*. Sometimes, the sample size for each stratum is dictated not by the relative size of the strata but by the allocation that yields the most precise estimates in the sense that standard errors of point estimates are minimal. In other words, in a sampling survey, we generally have allocated a fixed budget. This fixed budget includes fixed costs and variable costs that are similar for each stratum. Given a fixed budget, C, we want to select a sample of size n among different strata in such a way as to minimize the variance of the sample estimate. This kind of allocation of sample size is called *optimal allocation of sample*.

[5] The derivation of the sample size for proportional allocation defined in Eq. 20.13 and that of the sample size for optimal allocation with similar variable cost in each stratum, defined in Eq. 20.15, can be found in T. Yamane (1967), *Elementary Sampling Theory,* (Englewood Cliffs, NJ.: Prentice Hall), Chapter 6.

The optimal proportion of the sample that should be given to the jth stratum is

$$n_j = \frac{N_j s_j}{\sum_{j=1}^{H} N_i s_j} \cdot n \tag{20.14}$$

where S_j is the sample standard deviation for the jth stratum. For an optimal allocation of a stratified sample, the total sample n is given by the formula

$$n = \frac{1/N\left(\sum_{j=1}^{H} N_j s_i\right)^2}{N\sigma_{\bar{x}_{st}}^2 + 1/N\sum_{j=1}^{H} N_j s_j^2} \tag{20.15}$$

where $\sigma_{\bar{x}_{st}}^2$ is the desired variance for the sample mean.

Example 20.8 Sample Size for Stratified Random Sampling. Let's return to our hypothetical New York City law firm and use the information given in Example 20.3 to determine the sample size when the desired sample standard deviation, $\sigma_{\bar{x}_{st}}$, is \$1,900.

Substituting $N_1 = 475, N_2 = 50, N = 525,$ and $\sigma_{\bar{x}_{st}} = 1,900$ into Eq. 20.13, we can determine the sample size for a proportional allocation as

$$\begin{aligned} n &= \frac{475(11{,}620)^2 + 50(80{,}210)^2}{525(1{,}900)^2 + \frac{1}{525}\left[475(11{,}620)^2 + 50(80{,}210)^2\right]} \\ &= 146.69 \end{aligned}$$

For the degree of precision we desire, we should sample 147 of the lawyers in the firm if we plan to use a proportional allocation in our stratified sampling.

For an optimal allocation, the sample size is

$$\begin{aligned} n &= \frac{\frac{1}{525}[475(11{,}620) + 50(80{,}210)]^2}{525(1{,}900)^2 + \frac{1}{525}\left[475(11{,}620)^2 + 50(80{,}210)^2\right]} \\ &= 65.77 \end{aligned}$$

Obviously, using an optimal allocation in our stratified random sampling greatly reduces the size of our sample—from 147 to 66.

20.5 Two-Stage Cluster Sampling

When a researcher is interested in surveying a population that is dispersed throughout a large geographic region, neither simple nor stratified random sampling may be the best method for constructing the survey. Although a simple or stratified random sample still produces good estimates for determining population parameters, the expense of sampling across a large geographic region often dictates the need for alternative sampling techniques. Under these conditions, and also the cost when there are no reliable elements in the population to construct a sampling list, cluster sampling is preferred. This is because *two-stage cluster sampling* treats each cluster as a sampling unit.

In cluster sampling, we divide the population into clusters – geographically compact units such as congressional districts at the state level or political wards within a city. After dividing our population into clusters, we take a simple random sample of clusters and conduct a census in each of the sampled clusters. In other words, every individual in each of the sampled clusters is contacted. The advantage of cluster sampling over simple or stratified random sampling should be obvious. In conducting a census on a random sample of clusters, we can greatly reduce our costs by sampling in geographically compact areas. However, it should be noted that cluster sampling increases the sampling variance.

The technique used for cluster sampling parallels the approaches used in other sampling methods. To conduct a survey using the two-stage cluster sampling approach, we take the following steps[6]:

1. Divide the population into M clusters. For example, New York City might be divided into M voting districts.
2. Take a simple random sample of m sample clusters. Then take a simple random sample from each cluster. In the first stage, a random sample of m voting districts is selected. In other words, instead of selecting families one at a time, we have selected m groups of families, and, in our present case, each group of families lives in the same voting district. Then, in the second stage, random samples of $n_1, n_2, \ldots, n_m$ families are selected from each of the m districts' mth population observations, $N_1, N_2, \ldots, N_m$. Thus, our sample size $n = n_1 + n_2 + \ldots, + n_m$.
3. Compute an unbiased estimator of the population total:

$$\begin{aligned}\hat{X} &= \frac{M}{m}\sum_{i=1}^{m}\frac{N_t}{n_i}\sum_{i=1}^{n_i}x_{ij} \\ &= \frac{M}{m}\sum_{i=1}^{m}\hat{X}_i \end{aligned} \tag{20.16}$$

[6] See T. Yamane (1967), *Elementary Sampling Theory,* Chap. 8.

where

x_{ij} = the observation in the jth sample with sample size n_i and the ith cluster with sample size m
m = number of sampled clusters
N_i = number of population members in cluster i
M_i = number of population clusters

4. By dividing N into Eq. 20.16, obtain the estimated population mean $\hat{\mu}$ as

$$\hat{\mu} = \frac{\sum_{i=1}^{m} N_i \bar{x}_i}{(\bar{N})(m)} \tag{20.17}$$

where

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$$
$$\bar{N} = \frac{N}{M}$$

It can be shown that the variance associated with $\hat{\mu}$ is

$$\sigma_{\hat{\mu}}^2 = \mathrm{Var}(\hat{\mu}) = \frac{(M-m)}{Mm\bar{N}^2} \frac{\sum_{i=1}^{m} n_i^2 (\bar{x}_i - \bar{x})^2}{m-1} + \frac{1}{Mm\bar{N}^2} \sum_{i=1}^{m} N_i^2 \frac{(N_i - n_i)}{N_i} \frac{\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2}{(n_i - 1)} \tag{20.18}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{m} \bar{x}_i}{m}$$

If $n_i = N_i$, then Eq. 20.18 reduces to

$$\sigma_{\hat{\mu}}^2 = \mathrm{Var}(\hat{\mu}) = \frac{(M-m)}{Mm\bar{N}^2} \frac{\sum_{i=1}^{m} N_i^2 (\bar{x}_i - \bar{x})^2}{(m-1)} \tag{20.19}$$

If $\bar{N}$ is not available, we can use $\bar{n} = \sum_{i=1}^{m} n_i/m$ to substitute for $\bar{N}$ in both Eqs. 20.17 and 20.19.

Table 20.1 Sample family income data

Sample voting district, i	Mean income (thousands of dollars), $\bar{x}_i$	Number of families, N_i	(A) $N_i\,\bar{x}_i$	(B) $N_i^2(\bar{x}_i - \bar{x})^2$
1	30.62	30	918.6	89,956.80518
2	28.96	25	724	84,937.2736
3	21.56	28	603.68	284,742.6203
4	25.18	32	805.76	244,039.1616
5	33.56	27	906.12	36,311.28424
6	26.89	39	1,048.71	286,627.8896
7	24.56	40	982.4	412,554.4284
8	29.67	38	1,127.46	173,063.3216
9	40.12	31	1,243.72	237.9491353
10	53.16	33	1,754.28	171,312.5477
11	42.56	35	1,489.6	4,621.824256
12	56.37	37	2,085.69	339,701.0667
13	29.45	29	854.05	104,885.5586
14	50.66	40	2,026.4	161,359.6764
15	48.29	45	2,173.05	119,203.0865
16	42.39	43	1,822.77	5,808.451854
17	56.17	37	2,078.29	331,129.8125
18	45.89	29	1,330.81	23,378.28768
19	53.84	27	1,453.68	127,452.4272
20	49.26	28	1,379.28	58,557.80496
21	39.45	32	1,262.4	1,396.008714
22	42.57	45	1,915.65	7,719.028164
23	51.38	39	2,003.82	176,176.2949
24	55.66	29	1,614.14	190,296.2639
25	31.59	42	1,326.78	143,761.6989
Sum	1,009.8	860	34,931.14	3,579,230.573
$\bar{x} = 40.617604$				

5. Again, if the sample size is large, construct a $100(1 - \alpha)$ percent confidence interval:

$$\hat{\mu} - z_{\alpha/2}\sigma_{\hat{\mu}} < \mu < \hat{\mu} + z_{\alpha/2}\sigma_{\hat{\mu}} \tag{20.20}$$

Example 20.9 Population Mean Estimate for Average Family Income. A simple random sample of 25 voting districts is taken from a city with a total of 300 voting districts. Each family in the sample voting districts is surveyed to obtain information about family income. The sample data are listed in Table 20.1.

From the data of Table 20.1, we have $m = 25$ and $M = 300$. The total number of families in the sample is

$$\sum_{i=1}^{m} N_i = 30 + 25 + \ldots + 42 = 860$$

To obtain estimated average family income, we also need

$$\sum_{i=1}^{m} N_i \bar{x}_i = 918.6 + 724.0 + \ldots + 1,326.87 = 34{,}931.14$$

Substituting these figures into Eq. 20.17, we obtain

$$\hat{\mu} = \frac{\sum_{i=1}^{m} N_i \bar{x}_i}{\overline{N} m} = \frac{34{,}931.14}{860} = 40.6176$$

On the basis of the sample data, we estimate that annual family income is \$40,617.6.

In order to obtain an interval estimate, we need

$$\overline{N} = \frac{\sum_{i=1}^{m} N_t}{m} = \frac{860}{25} = 34.4$$

Also,

$$\frac{\sum_{i=1}^{m} N_i^2 (\bar{x}_i - \bar{x})^2}{m-1} = \frac{(30)^2(30.62 - 40.62)^2 + \ldots + (31.59)^2(31.59 - 40.62)^2}{24}$$
$$= 149{,}134.607$$

so

$$\mathrm{Var}(\hat{\mu}) = \frac{(275)(149{,}134.607)}{(300)(25)(34.4)^2}$$
$$= 4.6210$$

Taking square roots, we obtain $\sigma_{\hat{\mu}} = 2.1497$. For a 95 % confidence interval, $z_{\alpha/2}\ z_{.025} = 1.96$, so a 95 % confidence interval for the population mean is

$$40.6176 - (1.96)(2.1497) < \mu < 40.6176 + (1.96)(2.1497)$$

or

$$36.4042 < \mu < 44.8310$$

20.6 Ratio Estimates Versus Regression Estimates

So far in this chapter, we have been interested in making inferences about a population that are based on samples from that population. In this section, we consider two methods – the ratio method and the regression method – that can be used in conjunction with sampling to improve the parameter estimates based on sample data. (We discuss only the simple random sampling case, but both methods can also be used with stratified random sampling.) Note that the ratio method, which is easier than the regression method, is often used in sampling surveys.

20.6.1 Ratio Method

In the ratio method, an auxiliary variate y_i which is correlated with x_i is obtained for each unit in the sample. The population total Y of the y_i must be known. The value of y_i can be highly correlated with x_i (as are, say, sales and earnings), or y_i can be the value of x_i at some previous time when a complete census was taken.

To use the ratio method to estimate X, the population total of the x_i, we need to know Y, the population total of the y_i. The estimate of X, $\hat{X}_r$, is

$$\hat{X}_r = \frac{x}{y}Y = \frac{\bar{x}}{\bar{y}}Y \tag{20.21}$$

where

$y =$ sample total of y_i
$x =$ sample total of x_i
$\bar{y} =$ sample mean of $y_i = y/n$
$\bar{x} =$ sample mean of $x_i = x/n$
$Y =$ population total of y_i

If y is the value of x_i at some previous period, in the ratio method, we use the sample to estimate the relative change during the time interval.

Example 20.10 Ratio Estimate of Sales Prediction. Suppose a sales manager at Bono Corporation is interested in estimating the 1991 total sales at the company's 100 stores. To do this, he collects information for 1990 and 1991 sales from a simple random sample of 30 stores. The results of this sampling are presented in Table 20.2.

An estimate of 1991 sales can be produced in two ways. To use the ratio method, the sales manager needs to know overall Y (total sales in 1990). If 1990 total sales were \$14.3 million, then the ratio method yields the following prediction of 1991 sales:

Table 20.2 Sales for 1990 and 1991 (thousands of dollars)

Store	1990 sales	1991 sales
1	100.2	107.4
2	74.3	82.5
3	88.6	75.6
4	210.7	223.4
5	109.5	125.6
6	110.6	111.5
7	62.4	53.5
8	88.3	89.6
9	237.6	245.3
10	196.4	188.7
11	147.6	168.9
12	185.6	200.7
13	95.7	95.8
14	100.3	98.6
15	127.6	135.4
16	130.2	170.6
17	210.3	221.4
18	213.6	262.1
19	220.5	275.3
20	250.6	248.5
21	275.8	300.3
22	125.6	121.2
23	130.5	131.7
24	180.7	191.8
25	89.3	94.2
26	75.6	78.3
27	185.5	191.2
28	184.3	190.2
29	150.6	165.3
30	132.4	133.7
Mean	149.6966	159.2766
Ratio	1.063996	

$$\begin{aligned}\hat{X}_r &= (\bar{x}/\bar{y})Y \\ &= \frac{159.28}{149.70}(\$14{,}300{,}000) \\ &= \$15{,}215{,}124\end{aligned}$$

The second approach to estimating X is to use the sample mean per store to get the population total:

$$\begin{aligned}\hat{X}_r &= N\bar{x} \\ &= 100(159.28) = \$15{,}928{,}000\end{aligned}$$

This method doesn't utilize the 1990 sample information; hence, it is not as precise as the estimate that uses the 1990 sample information. This method, however, can be useful when 1990 sample information is not available.

20.6.2 Regression Method

A second approach to increasing the precision of estimates of population parameters based on sampling is the regression method. As we saw in Chaps. 13 and 14, simple linear regression enables us to relate two variables that are correlated with one another.

Suppose we are interested in μ_x, the population mean of x. To produce an estimate of μ_x, we use our knowledge of the fact that an auxiliary variable y_i is correlated with x_i. Again, we can employ simple random samples of y_i and x_i to produce $\bar{y}$ and $\bar{x}$, the sample means of y and x. In the regression method, the estimate of the population mean μ_x is

$$\hat{\mu}_x = \bar{x} + b(\mu_y - \bar{y}) \tag{20.22}$$

where

$\bar{y}$ = sample mean of y
μ_y = population mean of y (known)

Example 20.11 Regression Estimate of Sales Prediction. To illustrate the regression method, we will use the data given in Table 20.2. A simple linear regression of y_i (sales in 1991) and x_i (sales in 1990) is run. The results are

$$x_i = -7.45 + 1.11y_i$$

To use the regression method, we simply substitute the slope estimate $b = 1.11$ and $\mu_y = 14{,}300{,}000/100 = 143{,}000$ into Eq. 20.22:

$$\begin{aligned}\hat{\mu}_x &= \bar{x} + b(\mu_y - \bar{y}) \\ &= 159.28 + 1.11(143 - 149.70) \\ &= 151.84\end{aligned}$$

Again, we could produce an estimate of the population total as

$$\begin{aligned}X &= N\hat{\mu}_x \\ &= 100(151.84) \\ &= \$15{,}184{,}000\end{aligned}$$

20.6.3 Comparison of the Ratio and Regression Methods

You may have noticed some similarities between the regression method and the ratio method. Both methods use an auxiliary variate y_i, which is correlated with x_i. In fact, the ratio method is a special case of the regression method. These two methods are identical when the regression line passes through the origin.

Example 20.12 Labor Force Sampling. One important economic application of sampling is in determining structural changes in the labor force. The federal government conducts a census at the start of each decade. However, the Commerce Department's Bureau of the Census uses sampling to update census figures continually and to provide economists and other policy makers with labor force statistics such as the unemployment rate, employee wages, and the age, sex, and race of the labor force. The issue is how best to produce the labor force estimates.

The most widely used survey on the structure of the labor force is the current population survey (CPS). The CPS is a monthly survey that deals primarily with labor force data for the noninstitutional civilian population. Questions related to labor force participation are asked of each member in every sample household. In addition, supplementary questions regarding monetary income and work experience for the previous year are asked every March.

The present CPS sample was selected from the 1980 census files and consists of 60,000 occupied households. All 50 states and the District of Columbia are represented in the current CPS sample's 729 areas, which include 1,973 countries, independent cities, and minor civil divisions.

The estimates that the samples yield of the total noninstitutional civilian population of the United States by age, sex, and race are used to update census information. Through stratified sampling or two-stage cluster sampling and a technique such as the ratio method or the regression method, the Bureau of the Census is able to provide monthly estimates about the labor force.

20.7 Business and Economic Applications

Application 20.1 Sampling in an IRS Audit. We have discussed several ways of taking samples from a population. In this section, we apply these techniques to the accounting problem of auditing accounts receivable.

Suppose an Internal Revenue Service auditor is interested in determining whether the number of accounts receivable reported on LeClair Company's tax return is correct. Because LeClair has a total of 1,675 accounts receivable, the auditor has decided to sample the accounts receivable rather than to conduct a census.

Using a simple random sampling of 100 accounts receivable, the auditor finds a sample mean of \$127.84 and a sample standard deviation of \$42.62. LeClair has

reported that the mean value of its accounts receivable is $94.25. Should the auditor suspect that the company is underreporting its accounts receivable?

Because the auditor has information only on the sample mean, he decides to construct a 95 % confidence interval for the mean value of accounts receivable. If the mean value of accounts receivable reported by LeClair Company falls outside this interval, the auditor will consider this reason to investigate the company's tax return further.

To construct a 95 % confidence interval for the mean μ, the auditor uses $N = 1{,}675$, $n = 100$, $\bar{x} = \$127.84$, and $s = \$42.62$. He needs unbiased estimators for the population mean and the variance of $\bar{x}$. The sample mean $\bar{x}$ can be used as an unbiased estimator of the mean μ, and $\hat{\sigma}^2_{\bar{x}}$ can be calculated as follows:

$$\begin{aligned}\hat{\sigma}^2_{\bar{x}} &= \frac{s^2}{n} \cdot \frac{N-n}{N-1} \\ &= \frac{(42.62)^2}{100} \cdot \frac{1{,}675 - 100}{1{,}674} \\ &= 17.09 \\ \hat{\sigma}_{\bar{x}} &= \sqrt{17.09} = 4.13\end{aligned}$$

The 95 % confidence interval is

$$\bar{x} - Z_{\alpha/2}\hat{\sigma}_{\bar{x}} < \mu < \bar{x} + Z_{\alpha/2}\hat{\sigma}_{\bar{x}}$$

The value for $Z_{\alpha/2}$ can be found in Table A3 in Appendix A. It is $Z_{\alpha/2} = Z_{.025} = 1.96$, so

$$\begin{gathered}127.84 - (1.96)(4.13) < \mu < 127.84 + (1.96)(4.13) \\ 119.74 < \mu < 135.93\end{gathered}$$

Because the mean value of accounts receivable falls outside the 95 % confidence interval, the auditor's suspicions are aroused, prompting an investigation of LeClair Company's tax returns.

Application 20.2 Sampling Survey for 1977 Generic Drug Substitution Law in Wisconsin. In 1977, the state of Wisconsin passed a law that permitted the substitution of generic drugs for brand-name drugs when prescriptions were being filled.[7] Consumers had simply to request the substitution, and the pharmacist was legally bound to make it. The legislation was designed to save customers' money on their prescriptions. Thus, it proved disconcerting to the Department of Health and

[7] This application is drawn from G. A. Churchill, Jr. (1983), *Marketing Research: Methodological Foundations,* 3d ed., (Chicago: Dryden), pp. 441–442. Copyright© 1983 by The Dryden Press, reprinted by permission of the publisher.

Social Services when, some two and a half years after its passage, few customers were taking advantage of the law by asking for generic drugs.

Several hypotheses were advanced to explain why this was happening, including suggestions that the law had not been well publicized, that consumers were not aware of its existence, that consumers had unfavorable attitudes toward generic drugs, and that a number of personal and situational factors (such as age, income, household size, and education) were affecting customers' use of the law. It was decided that the best way to collect the needed information was through self-administered questionnaires. Because of the difficulty of obtaining an accurate mailing list, the questionnaires were to be delivered by hand but returned by mail. Further, investigators decided to determine the feasibility of the data collection and sampling plans by initially confining the study to Madison, the state capital, and to the main campus of the University of Wisconsin. They realized, of course, that restricting the original investigation in this manner would probably introduce some bias into the results because of a number of demographic differences between Madison and the remainder of the state.

The 1,000 households to be surveyed were selected in the following manner: First, detailed maps were used to divide the city into aldermanic districts. To ensure even geographic representation, samples were drawn from each aldermanic district. This was done by randomly choosing city blocks within each aldermanic district and then randomly selecting 10 households in each of the selected blocks to receive questionnaires.

Consider aldermanic district 11, for example, which was located on the city's west side. The study was to be limited to residents of Madison who were over 18. The total adult population in district 11 was 5,115, and the total adult population in the city was 122,016. The proportion of the total sample that was to come from aldermanic district 11 was thus $5{,}115/122.016 = .0419$. This meant that 42 households $[1{,}000(.0419) = 42]$ were to be included in the district 11 sample.

The 42 households were selected from 5 blocks within the district as follows: All blocks within the district were numbered. Then 5 blocks were randomly selected from this larger set of blocks. On each of the first 4 blocks that were selected, 10 households were interviewed, and 2 households were interviewed on the fifth block. The households were selected by first going around the block to count the number of dwelling units. Each field worker was to begin the count at the southwest corner, following the detailed instructions that were provided. Suppose, for example, that there were 50 dwelling units in a selected block. The field worker was then instructed to generate a random start between 1 and 5, using the table of random numbers each carried. If the number was, say, 2, the field worker was to drop off questionnaires at the second, seventh, and twelfth households, and so on in the initial numbering scheme.

Application 20.3 Acceptance Quality Sampling for Quality Inspection. The acceptance sampling used in inspection which has been discussed in Chaps. 10 and 11 can use either simple or stratified random sampling technique to select the sample.

Suppose a lot consists of 44,000 items from four different machines:

Machine	Lot size (strata size)	Size
1	20,000	250
2	6,000	70
3	8,000	85
4	10,000	150
	44,000	555

The acceptance or rejection decision could be based on the entire lot of 44,000 units using a single sample of 555 units drawn randomly from the entire lot.

By stratifying the items and basing the acceptance or rejection decision on the quality of the lot from each machine, better information is obtained since the number produced by each machine varies significantly. If quality differences exist between machines, stratified sampling can be used to discover this fact.

20.8 Summary

In this chapter, we explained why sampling surveys are needed for analyzing business and economic data. Then, we looked at three sampling methods: simple random sampling, stratified random sampling, and two-stage cluster sampling. Next we investigated how ratio and regression methods are used to estimate the total value and the mean value of a population on the basis of sample information. Finally, we explored some applications of sampling surveys.

Questions and Problems

1. Under what conditions might stratified random sampling or cluster sampling be preferable to simple random sampling? Explain.
2. Suppose we are interested in the proportion of college seniors in a class of 900 who will be attending graduate school. A survey of 50 seniors reveals that 15 will be attending graduate school. We want to estimate the population proportion p, given this information: $N = 900$, $n = 50$, $\alpha = 10\ \%$.
3. Suppose you are a financial consultant trying to determine whether a group of 1,500 country club members represents a good source of potential clients for a real estate firm in New Jersey. To determine the potential business, you decide to analyze the size of the club members' purchases of homes. A random sample of 90 members produces a sample mean of \$280,000 with a sample standard deviation of \$75,000. Using this information, construct a 95 % confidence interval for the mean purchases $N = 1{,}500$, $n = 90$, $\bar{X} = \$280{,}000$, and $s = \$75{,}000$.
4. If X is a normal variable with known variance equal to 650, how large a sample must we take to be 90 % confident that the sample mean will not differ from the true mean by more than $\pm$ 5 units?

5. If a normal population is known to have σ equal to 10, how large a sample should we take in order to be 90 % confident that the sample mean will not differ from the population mean by more than $\pm$ .75 units?
6. A sales manager wants to know what proportion of her accounts are inactive. How many accounts should she examine if she wants her confidence interval to be no more than $w = .08$ (w being the desired maximum width of the confidence interval) with 95 % confidence.
7. A company manager wishes to estimate the mean length of time μ it takes company crews to do certain jobs. She wants to estimate μ within $\pm$ 7 min with 95 % confidence. Because the value of σ, the population standard duration, is unknown, she took a preliminary sample of $n = 20$ jobs and found that the 20 job completion times had a standard deviation of $s = 18$ min. How much larger should she make her sample to obtain the desired confidence interval?
8. The claims manager for an insurance company would like to know the mean amount of automobile insurance repair claims paid by his company. He took a sample of $n = 25$ claims and found $\bar{x} = \$950$ and $s = \$280$. How much larger should his sample be if he wants to estimate the mean payment to within $\pm$ 50 with 90 % confidence?
9. A marketing manager is interested in the number of trips per month that people take to a nearby shopping center. Denote the number of trips per month by X. The manager feels that the variance of X is 16 for women and 9 for men. He decides to stratify by sex, and there are 100 males and 200 females in the population of interest. He takes a random sample of 10 males and another random sample of 15 females and gets the following results:

Males	4	2	1	1	4	7	10	5	7	3					
Females	6	9	12	4	2	10	9	7	5	8	16	14	10	8	13

   (a) Estimate the mean of each of the two strata and the mean of the entire population by pooling the samples.
   (b) Determine the variance of the pooled samples.

10. Is the plan in question 9 stratified sampling with proportional allocation? When might proportional allocation *not* be the best form of allocation in a stratified sampling plan?
11. The number of cars licensed in a particular state last year was 4.8 million; the number of cars licensed in a neighboring state was 3.9 million. For the current year, the officials of the neighboring state estimate that there will be 4.65 million cars registered. Can you use this information to estimate the number of cars that will be registered in the first state in the current year?
12. In questions 9, suppose the marketing manager is interested in the *total* number of trips to the shopping center in a given month, not in the average number of trips per person for the month.

(a) Estimate the population totals for the two strata, and determine the sample variance of estimation in each case.
(b) Estimate the total for the entire population of males and females, and determine the variance of estimation.

13. Outline the basic steps used in designing a sampling study. What problems may you encounter if your study is poorly planned?
14. What is simple random sampling? What are the advantages and disadvantages of this technique compared to other sampling techniques?
15. What is sampling error? What is a nonsampling error? Give some examples of each.
16. What is two-stage cluster sampling? What are the benefits and disadvantages of this technique compared to other sampling techniques?
17. Suppose you are working for a political consulting firm that is trying to forecast the outcome of a presidential election. Because of the time and cost involved in conducting a simple random sample across all 50 states, you've been asked to devise a sampling strategy to predict the outcome of the election. Given the time and money constraints, what sampling technique will you propose?
18. Use the information given in Example 20.1 to determine the size of the sample if you want a level of precision of $\hat{\sigma}_{\bar{x}} = \$12{,}000$.
19. Again using the data from Example 20.1 and question 18, construct a 95 % confidence interval.
20. Use the data in Example 20.2 to determine the size of the sample you need if you want a level of precision of .05.
21. Suppose you have decided to conduct a sampling survey on the salaries of baseball players. There are essentially two types of players, those not eligible for free agency and those eligible for free agency. If you believe that the salaries of players who are eligible for free agency will differ from those of players who are not, what type of sampling method should you use?
22. Explain how we can use the ratio method to improve the parameter estimates based on sample data.
23. Explain how we can use the regression method to improve the parameter estimates based on sample data.
24. What is the jackknife method? What advantages does it offer? Briefly explain how we use the jackknife method.
25. Suppose you have decided to conduct a study to determine whether accounting majors should be required to take statistics. Briefly explain how you would set up this study. What problems might you encounter? How would you deal with these problems?
26. Using the business section of the *Wall Street Journal* or the *New York Times,* obtain a list of all stocks traded on the American Stock Exchange. Use a random sample of 15 stocks to compute the mean percentage increase in the prices of these stocks over the last month. Compare your result to the actual change in the AMEX index over that period.

27. The citizens for Fair Taxes are interested in the average property tax paid by the 2,000 residents of their city. A random sample of 50 of these households had a mean property tax of $1,472 with a standard deviation of $311

    (a) Find an estimate of the variance of the sample mean.
    (b) Find a 95 % confidence interval for the population.

28. The Mom and Pop Grocery Store has 115 employees. In a random sample of 30 of these employees, the mean number of days that an employee was late each year was 14 days, and the sample standard deviation was 3.4 days. Find a 99 % confidence interval for the mean number of days late each year.
29. The academic advisor to the football team of Rah Rah University is interested in the mean number of hours that players spend studying during the football season. Of the 100 members of the football team, 35 were randomly sampled and found to study an average of 22.5 h per week with a standard deviation of 8.1 h. Find a 90 % confidence interval for the mean number of hours that the football players study.
30. Suppose a quality control expert is interested in drawing a random sample of 100 light bulbs from a case of 10,000. Explain how a table of random numbers could be used to do this.
31. Suppose the quality control expert of question 30 knows from past experience that the number of defective light bulbs in a case of 10,000 has a population standard deviation of 221. She would like to compute a 99 % confidence interval for the population mean with a precision of 50. How many light bulbs should she sample?
32. A city is required to report to the state the mean property tax of its citizens. From previous years, officials know that the population standard deviation is likely to be $975. A 99 % confidence interval is desired with a precision of $500. How many of the city's 1,800 households should be sampled?
33. An advertising executive is interested in how viewers looked upon a television ad by McDonald's in New York City (either favorably or unfavorably). Briefly explain how the executive could analyze this question by using sampling. Is one method of sampling preferable to another?
34. In order to correctly assess property taxes in the state, New Jersey has decided to require that all municipalities report the average home price in their districts. From past years, one municipality estimates that the population standard deviation for its 2.500 homes is $51,721. If the town would like to produce a 95 % confidence interval with a level of precision of $10,000, how many homes should be sampled?
35. The Students for an Affordable Education have asked you to estimate the average amount of money spent per semester on textbooks. To produce this estimate, you have decided to randomly sample 25 members of your Introduction to Economics course. Are there any problems associated with this sample? What other sampling techniques might you use?
36. An auditor would like to estimate the total value of a corporation's accounts receivable. From previous years, the auditor has found the population standard

deviation to be \$125 for the 1,000 accounts receivable. If the auditor would like to have a level of precision of \$100, with a 95 % confidence interval, how large a sample should he select?

37. Suppose the auditor in question 36 decides to divide the accounts receivable into strata. He would like a desired standard deviation of \$10. Determine the total number of sample observations under (a) proportional allocation and (b) optimal allocation.

Stratum	Population size	Estimated standard deviation
1	250	\$85
2	325	\$125
3	225	\$50
4	200	\$100

38. A first-year chemistry class consists of 150 students. A random sample of 50 of these students reveals that 31 are majoring in engineering. Find a 95 % confidence interval for the proportion of students in this class who are majoring in engineering.
39. Explain whether the confidence interval gets wider or narrower when

    (a) The confidence interval is 99 % instead of 95 %.
    (b) The number of observations in the sample decreases from 100 to 50.
    (c) The population standard deviation is smaller.

40. The student government of your school has asked you to survey students in order to determine how many hours the library should be open. You have decided to conduct a stratified sample by class year: first-year, sophomore, etc. What factors must you account for in determining the number of sample observations in each stratum?
41. A movie studio executive wants to poll a sample of movie goers to determine how viewers will react to a new movie. Briefly explain how the sample should be designed.
42. A quality control engineer at the National Bullet Company wants to test for the number of defective bullets (duds) in a case of 1,000. From past experience, he knows that population standard deviation per box is 20 bullets. If he would like to estimate the mean number of duds with a level of precision of 5 bullets, with a 95 % confidence interval, how large a sample should he take?
43. Suppose you want to estimate a population mean μ. From your sample, you find that the sample mean is 200, the sample standard deviation is 30, the total population is 500, and the sample drawn is 30. Find a 90 % confidence interval for the population mean.
44. Suppose you want to estimate a population proportion p. The total population consists of 1,000, your population consists of 100, and the sample proportion is .42. Find a 95 % confidence interval for the population proportion.

45. A survey based on a stratified sample produced the following information:

Stratum				
	1	2	3	4
N	1,000	2,000	2,200	3,200
n	40	50	42	75
$\bar{x}\ X$	27.6	30.4	18.7	32.5
s^2	4.2	7.1	6.4	2.8
$\hat{p}\ P$	.5	.6	.3	.6

where $\bar{x}$, s^2, and $\hat{p}$ are the sample mean, sample variance, and sample proportion, respectively.

(a) Find a 90 % confidence interval for the population mean.
(b) Find a 95 % confidence interval for the population proportion.

46. The results of a sample survey based on cluster sampling follow:

Cluster					
	1	2	3	4	5
n	3	7	8	2	9
x_i	6.2	8.3	5.4	7.3	9.1

$M = 2{,}500$ and $N = 100$
Find a 90 % confidence interval for the population mean.

47. A survey based on a stratified sample produced the following information:

Stratum					
	1	2	3	4	5
N	1,900	3,000	3.500	4,200	4,100
n	30	51	72	65	83
$\bar{x}\ X$	22.6	40.5	28.7	22.5	25.4
s^2	5.2	4.1	7.4	3.8	5.1
$\hat{p}\ p$	0.3	0.1	0.2	0.3	0.4

where $\bar{x}$, s^2, and $\hat{p}$ are the sample mean, sample variance, and sample proportion, respectively.

(a) Find a 90 % confidence interval for the population mean.
(b) Find a 95 % confidence interval for the population proportion.

48. The results of a sample survey based on cluster sampling follow:

Cluster					
	1	2	3	4	5
n	2	5	7	3	8
x_i	4.2	2.3	8.4	5.3	6.1

$M = 5{,}500$ and $N = 300$
Find a 95 % confidence interval for the population mean.

49. A supermarket manager wants to know what type of people purchases health foods and wishes to determine why segments of the population resist buying these products. Discuss how you would go about setting up a study to provide this information. What difficulties might you encounter?
50. A real estate developer wants to determine which features of a house have been most influential in determining its selling price. Describe how you would set up a study to provide this information. What difficulties might you encounter?
51. Suppose the *American Economic Review* is interested in knowing whether student subscribers and regular subscribers differ in their viewpoints on the articles offered in the journal. Explain how you would set up such a study.
52. A city consists of a total of two million residents and is divided into three boroughs that have 750,000, 900,000 and 350,000 residents, respectively. The city council is considering building a new baseball stadium. If the project is undertaken, it will be financed by an increase in taxes. In order to determine how the city feels about the new stadium, independent random sampler of 500 adults from each borough were taken. The numbers in favor of the stadium were found to be 325, 201, and 400, respectively.

    (a) Using an unbiased estimation procedure, find an estimate of all adults in the city who favor the stadium.
    (b) Find a 90 % confidence interval for this population proportion.

53. Suppose a large Wall Street law firm has 500 lawyers of whom 95 are partners and 405 are associates. A random sample of 15 partners finds that 11 own their own homes, and a random sample of 25 associates finds that 15 own their own homes.

    (a) Find an estimate of the proportion of all lawyers in this firm who own their own homes, using an unbiased estimation procedure.
    (b) Find a 95 % confidence interval for all lawyers in this firm who own their own homes.

54. Refer to question 53. Suppose a random sample of the 12 partners reveals that 6 of them graduated from Ivy League schools and a random sample of 20 associates finds that 14 graduated from Ivy League schools:

    (a) Find an estimate of the proportion of all lawyers in this firm who graduated from Ivy League schools.
    (b) Construct a 99 % confidence interval for all lawyers in this firm who graduated from Ivy League schools.

55. The president of a local union is interested in the mean value of bonuses awarded to a company's employees. The company has 35 divisions, and a simple random sample of 5 of these is taken. The following table gives the results of this sample:

Division sampled	Number of employees	Mean bonus
1	55	$ 70
2	97	101
3	60	40
4	35	89
5	72	56

(a) Find a point estimate of the population mean bonus per employee.
(b) Find a 90 % confidence interval for the population mean.

56. A personnel manager is interested in the average age of the company's 872 employees. Suppose he takes a simple random sample of 35 of these employees and finds the sample standard deviation of their ages to be 12.3 years. The personnel manager wants to obtain a 95 % confidence interval for the population mean age with a level of precision of 2.4 years on each side of the sample mean. How many sample observations must he take?
57. A company has a fleet of 322 automobiles. A random sample of 35 of the cars finds that the sample standard deviation of annual repair costs is $272. Company planners want to construct, for the overall mean of annual repair costs, a 90 % confidence interval that extends $100 on either side of the sample mean. How many additional sample observations must they take?
58. Mention some situations in which two-stage cluster sampling should be used.
59. A clothing store has an inventory of 920 different items. In order to estimate the total dollar value of inventories, an auditor takes a simple random sample of the items. On the basis of last year's data, the population standard deviation is estimated to be $97. The auditor would like to produce, for the population total, a 95 % confidence interval that extends $16.30 on each side of the sample estimate. How large a sample size is necessary to meet this requirement?
60. Suppose a corporation is interested in the proportion of employees who favor a new child care program. The corporation has 750 employees from which it wants to take a simple random sample. The planners would like to make the sample large enough so that they can produce a 90 % confidence interval that extends no more than 7 % on each side of the sample proportion in favor of the new program. How large a sample should they take? Assume that the sample standard deviation is .24.
61. A movie theater chain has 35 theaters in California, 50 in New York, and 45 in Pennsylvania. Management is considering adding a new snack item to its concession stands. In order to determine whether this new snack will be a success, management tested the product in 10 theaters in California, 12 in New York, and 9 in Pennsylvania for 1 month. The sample means and standard deviations for the numbers of purchases are shown here.

	California	New York	Pennsylvania
Mean	100.4	98.2	77.6
Standard deviation	40.3	21.7	45.1

(a) Use an unbiased estimation procedure to find an estimate of the mean number of purchases per movie theater in a month for all 130 movie theaters.
(b) Find an estimate of the variance of the estimator in part (a), using an unbiased estimation procedure.
(c) Find a 90 % confidence interval for the population mean number of purchases per theater.

62. Suppose that in question 61 we are interested in knowing how large a sample to take for:

(a) A proportional allocation.
(b) An optimal allocation, assuming the stratum population standard deviations are the same as the corresponding sample values.

63. A delivery company has a fleet of 502 trucks. A random sample of 35 of these trucks finds that the sample standard deviation of annual repair costs is \$753. If you would like to construct, for the overall mean of annual repair costs, a 95 % confidence interval that extends \$250 on either side of the sample mean, how many additional sample observations must you take?
64. A fast-food chain has 25 restaurants in Alabama (Ala.), 20 in Louisiana (La.), 25 in Texas (Tex.), and 32 in Arkansas (Ark.). Management is considering adding a hamburger item to its menu. In order to determine whether this burger will be a success, management tested the product in 15 restaurants in Alabama, 11 in Louisiana, 18 in Texas, and 5 in Arkansas for 1 week. The sample means and standard deviations for the numbers of purchases are shown here.

	Ala.	La.	Tex.	Ark.
$\bar{x}$	127.5	221.3	99.7	127.6
s	83.3	43.8	27.6	70.2

where $\bar{x}$ and s are the mean and standard deviation, respectively.

(a) Use an unbiased estimation procedure to find an estimate of the mean number of purchases of the burgers for all 102 restaurants
(b) Find an estimate of the variance of the estimator in part (a), using an unbiased estimation procedure.
(c) Find a 99 % confidence interval for the population mean number of burgers sold per restaurant.

65. Suppose that in question 64, a sample of 35 restaurants is to be taken. Determine how the sample should be allocated among the three states for

(a) A proportional allocation.
(b) An optimal allocation, assuming the stratum population standard deviations are the same as the corresponding sample values.

66. A company that operates three different types of factories is interested in the number of defective products produced. The following table gives the results of a sampling study done on the number of defective parts:

Number of defective parts in factories of			
	Type 1	Type 2	Type 3
N_i	75	90	100
n_i	10	15	20
$\bar{x}_i$	12.3	11.5	16.4
s_i	4.7	7.5	3.4

(a) Find an estimate of the total number of defective parts, using an unbiased estimation procedure.
(b) Find a 95 % confidence interval for this total.

67. Use the information given in question 66 to find an estimate of the mean number of defective parts, using an unbiased estimation procedure. Also find a 99 % confidence interval for this value.
68. Refer to question 66. Suppose a sample of 30 factories is to be taken. Determine how many factories of each type the company should select when it is using

(a) A proportional allocation.
(b) An optimal allocation, assuming the stratum population standard deviations are identical to the corresponding sample values.

69. An auditor is interested in estimating the population mean value for Aloha Company's accounts receivable. The population has been divided into three strata. The accompanying table gives information on the strata and the estimated standard deviations.

Stratum	Population size	Estimated standard deviation
A	600	\$175
B	1,005	220
C	700	195

$\sigma_{\bar{x}} = 12.75$. Assume you would like a 95 % confidence interval to extend \$25 on each side of the estimate.

(a) Determine the total sample size for a proportional allocation.
(b) Determine the sample size for an optimal allocation.

70. A company is interested in estimating the value of accounts receivable for its 75 stores. To do this, the company collects information on 1990 and 1991 accounts receivable from a simple random sample of 10 stores. The results of this sampling study are given in the table.

Store	1990 accounts receivable	1991 accounts receivable
1	$15,600	$16,200
2	9,510	8,900
3	27,000	29,000
4	18,000	19,200
5	32,000	32,500
6	22,200	19,000
7	25,200	28,500
8	15,000	17,000
9	19,600	24,000
10	9,900	11,100

Suppose 1990 accounts receivable were $ 1.4 million. Use the ratio method to estimate the accounts receivable for 1991.

71. Refer to question 70. Use the MINITAB program in terms of the regression approach to forecast accounts receivable in 1991. Compare your results to those you got with the ratio method in question 70.
72. A company is interested in estimating sales for its 200 stores. To do this, the company collects information on 1990 and 1991 sales from a simple random sample of 12 stores. The results of this sampling study are given in the table.

Store	1990 sales	1991 sales
1	$155,600	$160,200
2	89,540	98,900
3	275,400	282,000
4	180,900	190,200
5	325,000	320,500
6	222,200	199,900
7	250,400	278,400
8	151,000	178,000
9	190,600	240,000
10	99,900	115,100
11	311,000	354,000
12	272,500	295,000

Suppose 1990 sales were $27 million. Use the ratio method to estimate the sales for 1991.

73. Refer to question 72. Use the MINITAB program in terms of the regression approach to forecast sales in 1991. Compare your results to those you got with the ratio method in question 70.

74. An auditor is interested in estimating a company's bad accounts. He collects information on bad accounts for the company's 200 stores using a random sample of 25 stores. The results are given in the table.

	1990 bad accounts	1991 bad accounts
Mean	$127,000	$135,000

The total amount of bad accounts in 1990 was $24 million. Use the ratio method to forecast the bad accounts in 1991.

75. Refer to question 70. Suppose the company would like to construct a 90 % confidence interval for the increase in accounts receivable. Use the jackknife method to construct this confidence interval.
76. Refer to question 72. Use the jackknife method to construct a 95 % confidence interval for the increase in sales.
77. Suppose a corporation is interested in the proportion of employees who favor a new "flex time" work schedule. The corporation has 420 employees from which the planners wish to take a simple random sample. The planners would like to make the sample large enough so that they can produce a 95 % confidence interval that extends no more than 5 % on each side of the sample proportion in favor of the new program. How large a sample should they take?
78. A company operating three factories that use different types of production is interested in the number of defective products produced. The following table gives the results of a sampling study done on the number of defective parts.

Number of defective parts in production of			
	Type 1	Type 2	Type 3
Ni	55		150
ni	12	9	40
xi	22.3	35.4	26.3
si	10.7	9.5	13.4

(a) Find an estimate of the total number of defective parts. Using an unbiased estimation procedure.
(b) Find a 90 % confidence interval for this total.

79. Use the information given in question 78 to find an estimate of the mean number of defective parts, using an unbiased estimation procedure. Also find a 99 % confidence interval for this value.
80. Refer to question 78. Suppose a sample of 50 factories is to be taken. Determine how many factories of each type the company should select when it is using

(a) A proportional allocation
(b) An optimal allocation, assuming the stratum population standard deviations are identical to the corresponding sample values

81. A quality control engineer at the Brite Lite Light Bulb Company wants to test for the number of defective light bulbs in a case of 1,200. From past experience, he knows that the population standard deviation per box is 25 bulbs. He would like to estimate the mean number of defective bulbs with a standard deviation of 8 bulbs. How large a sample should he take?
82. A market research group takes a random sample of 5 of a city's 41 voting districts. Each household in each sampled district is questioned on the number of hours its members watch television per day. The results of the sample follow:

District	Number of households	Mean number of hours of TV per day
1	53	3
2	31	5
3	17	6
4	28	4
5	41	7

Find a 90 % confidence interval for the population mean number of hours of television watched.

83. Suppose you want to estimate the proportion of a population of voters. A random sample reveals that the sample proportion is .42; $N = 2{,}500$ and $n = 500$. Find a 99 % confidence interval for the population proportion.
84. The results of a sample survey based on cluster sampling are

Cluster					
	1	2	3	4	5
n	3	2	8	9	6
x_i	50.2	42.3	38.4	51.3	36.1
$M = 4{,}000$	and $N = 250$				

Find a 90 % confidence interval for the population mean.

85. You have been hired to design a sampling procedure for testing a new antiacne drug. One hundred people will be used in the study: half will receive the new drug, and half will receive the standard acne drug. Explain how you would decide which people get which drug. Be specific.
86. Professor Anderson needs a sample of students from his university with 13,200 students. He uses a table of random numbers to select 130 numbers between 1 and 13.200. His sample is a.

    (a) Simple random sample
    (b) Systematic sample
    (c) Stratified sample
    (d) Cluster sample

87. Professor Anderson's university has 13,200 students, of which are 7,020 are undergraduates, 4,200 are master students, and 1,800 are Ph.D. students.

He decides to randomly select 70 undergraduates, 42 master students, and 18 Ph.D. students. His sample is a.

(a) Simple random sample
(b) Convenient sample
(c) Stratified sample
(d) Cluster sampling

88. The president of Prof. Anderson's university is interested in the proportion of students favoring the new campus bus system.

(a) Since there are 13,200 students and the president would like to have a 95 % confidence interval to extend 3 % on each side of the mean proportion, how many students should be surveyed?
(b) Suppose a sample with the sample size being determined by (a) and 516 of them favoring the new campus bus system. Construct a 95 % confidence interval for the population proportion.

89. Prof. Anderson would like to estimate the monthly expenditure on mobile phone communication. If he would like a level of precision of $5, determine the total number of sample observations under a proportional allocation.

Stratum	Students	Estimated standard deviation
Undergraduate	7,200	$35
Master	4,200	$25
Ph.D.	1,800	$15

90. Use the data in Problem **89** to determine the appropriate sample sized under an optimal allocation rule.

Appendix 1: The Jackknife Method for Removing Bias from a Sample Estimate

In this appendix, we discuss the jackknife method, which can be used in conjunction with sampling to remove the bias of an estimator and to produce confidence intervals.

The jackknife is a general technique that can be applied to any linear estimator. It works by using the original sample to create a new set of "pseudovalues." The jackknife procedure involves the following steps:

1. The n sample values are divided into m subsets, and m is set equal to n in many applications. For example, removing one piece of data at a time leaves $m = n$ subsets of data with $(n - 1)$ observations in each set.
2. An estimate based on all the data is calculated. Call this value x_{All}.

3. An estimate based on all the data except the data from the first of the m subsets is calculated; call it x_{-1}. Estimates of x_{-2}, x_{-3}, . . . , x_{-m} are also calculated.
4. The "pseudovalue" x_1 is calculated as

$$x_1 = x_{\text{All}} + (m-1)(x_{\text{All}} - x_{-1}) \tag{20.23}$$

Likewise, we can calculate x_2, x_3,. . ., x_m. These pseudovalues will constitute a "pseudosample" that acts like a random sample. Alternatively, Eq. 20.23 can be rewritten as

$$x_1 = mx_{\text{All}} - (m-1)x_{-1} \tag{20.24}$$

5. The mean $\bar{x}$ and the standard deviation s of the pseudosample can now be calculated and used to produce confidence intervals. For example, a 90 % confidence interval for the population mean μ can be defined as

$$\mu = \bar{x} \pm t_{.05}\frac{s}{\sqrt{m}} \tag{20.25}$$

where $t_{.05}$ is the t statistic with the significance level $\alpha = .05$.

It may not be clear from an introduction of the jackknife technique why this procedure is preferable to a simple or a stratified random sample. It has been shown that when the original estimate is biased but is asymptotically unbiased – that is, unbiased in large samples – jackknifing often eliminates the bias. Also, the jackknife procedure makes it possible to compute confidence intervals for the population parameters when the samples taken are small and the population standard deviation is unknown.

Example 20.13 Removing the Bias of Accounts Receivable Estimates. Suppose an auditor is interested in determining the mean growth rate of uncollectible accounts receivable. This figure will help the auditor find out whether a store has an abnormally high increase in uncollectibles.

To obtain these estimates, the auditor randomly samples the uncollectibles of six department stores in 1989 and 1990. The results of this sample are given in Table 20.3. The auditor is interested in constructing a 95 % confidence interval for the ratio in uncollectibles. In Table 20.4, we present the ratio in uncollectibles, u, where

$$u = \frac{1990\,\text{uncollectibles}}{1989\,\text{uncollectibles}}$$

The issue now before us is how we compute a confidence interval for u by using the jackknife method.

Table 20.3 Random sample of uncollectibles for six stores

Store	1989 uncollectibles	1990 uncollectibles
AAA	$200,000	$225,000
BBB	$84,000	$92,000
CCC	$127,000	$152,000
DDD	$12,000	$13,500
EEE	$375,000	$390,000
FFF	$27.000	$42,000
Total	$825,000	$914,500

Table 20.4 *U* Ratios for six different stores

Store	*U*
AAA	1.125
BBB	1.095
CCC	1.197
DDD	1.125
EEE	1.040
FFF	1.556

The first step in the jackknife procedure is to calculate x_{All} the observation based on all the data. From our example, this is the ratio of total uncollectibles in 1990 to total uncollectibles in 1989.

$$x_{\text{All}} = \frac{914{,}500}{825{,}000} = 1.108$$

Next we compute x_{-1}, using all the data except the data of AAA Company.

$$x_{-1} = \frac{914{,}500 - 225{,}000}{825{,}000 - 200{,}000} = \frac{689{,}500}{625{,}000} = 1.103$$

Similarly, we compute x_{-2} by deleting the data of BBB Company, x_{-3}, by deleting the data of CCC Company, and so on.

$$x_{-2} = \frac{914{,}500 - 92{,}000}{825{,}000 - 84{,}000} = 1.110$$
$$x_{-3} = \frac{914{,}500 - 152{,}000}{325{,}000 - 127{,}000} = 1.092$$
$$x_{-4} = \frac{914{,}500 - 13{,}500}{825{,}000 - 12{,}000} = 1.108$$
$$x_{-5} = \frac{914{,}500 - 390{,}000}{825{,}000 - 375{,}000} = 1.166$$
$$x_{-6} = \frac{914{,}500 - 42{,}000}{825{,}000 - 27{,}000} = 1.093$$

Using this information, we calculate our pseudovalues $x_1, x_2, \ldots, x_6$ in accordance with Eq. 20.23.

$$\begin{aligned}
x_1 &= x_{\text{All}} + (m-1)(x_{\text{All}} - x_{-1}) \\
&= 1.108 + 5(1.108 - 1.103) \\
&= 1.133 \\
x_2 &= 1.108 + 5(1.108 - 1.110) \\
&= 1.098 \\
x_3 &= 1.108 + 5(1.108 - 1.092) \\
&= 1.188 \\
x_4 &= 1.108 + 5(1.108 - 1.108) \\
&= 1.108 \\
x_5 &= 1.108 + 5(1.108 - 1.166) \\
&= .818 \\
x_6 &= 1.108 + 5(1.108 - 1.093) \\
&= 1.183
\end{aligned}$$

We can now use these pseudovalues, $x_1, \ldots, x_6$, as our pseudosample to construct our confidence interval. First, we compute the sample mean $\bar{x}$ of the pseudosample. Then we compute the sample standard deviation s of the pseudosample. Finally, we construct a confidence interval, using Student's t distribution (it is given in Table A4 of Appendix A).

$$\begin{aligned}
x &= \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6}{6} \\
&= \frac{6.528}{6} \\
&= 1.088 \\
s &= \left\{ \frac{1}{n-1} (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_6 - \bar{x})^2 \right\}^{1/2} \\
&= \left\{ \frac{1}{5} [(1.133 - 1.088)^2 + (1.098 - 1.088)^2 + (1.188 - 1.088)^2 \right. \\
&\quad \left. + (1.108 - 1.088)^2 + (.818 - 1.088)^2 + (1.183 - 1.088)^2] \right\}^{1/2} = .137
\end{aligned}$$

For a 95 % confidence interval, we use $t_{.025}$ with 5 degrees of freedom.

$$t_{.025.5} = 2.57$$

Then the 95 % confidence interval in terms of Eq. 20.25 can be computed as

$$\bar{x} - t_{.025}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{.025}\frac{s}{\sqrt{n}}$$
$$1.088 - 2.57(.137/\sqrt{6}) < \mu < 1.088 + 2.57(.137/\sqrt{6})$$
$$.944 < \mu < 1.232$$

The 95 % confidence interval for the increase in uncollectibles is the interval between 0.944 and 1.232.

Chapter 21
Statistical Decision Theory: Methods and Applications

Chapter Outline

Key Terms

Statistical decision theory	Worst outcome
Bayesian decision statistics	Expected monetary value
Decision theory	Utility analysis
Actions	Total utility
Alternatives	Marginal utility
States of nature	Utility function
Event	Risk-averse
Outcomes	Risk-neutral
Payoff	Risk lover
Probability	Expected utility
Prior probability	Decision node
Posterior probability	Event node
Maximin criterion	Systematic risk

(continued)

C.-F. Lee et al., *Statistics for Business and Financial Economics,*
DOI 10.1007/978-1-4614-5897-5_21, © Springer Science+Business Media New York 2013

Statistical decision theory	Worst outcome
Minimax regret criterion	Market risk
Capital market line	Market portfolio
Sharpe investment performance measure	Lending portfolio
	Borrowing portfolio
Capital asset pricing model	Market risk premium
Treynor investment performance measure	Present value
	Net present value
Statistical distribution method	Jensen investment performance measure

21.1 Introduction

In business, decision making is at the heart of management. Using statistics as a guide, this chapter introduces and examines decision making in business and economics in terms of statistical decision theory. The branch of statistics called *statistical decision theory* is sometimes termed *Bayesian decision statistics*, in honor of research presented over 200 years ago by the English philosopher the Reverend Thomas Bayes (1702–1761). Nevertheless, statistical decision theory is a new branch of statistics. Propelled by research by Howard Raiffa, John Pratt, and Leonard Savage (among others), it developed rapidly in the 1950s, and it now occupies an important place in statistical literature. In contrast to classical statistics, where the focus is on estimation, constructing intervals, and hypothesis testing, statistical decision theory focuses on the process of making a decision. In other words, it is concerned with the situation in which an individual, group, or corporation has several feasible alternative courses of action in an uncertain environment.

This chapter discusses methods for selecting the best management alternatives by using statistical decision theory. Here are a few examples of statistical decision problems associated with business decision making:

1. Manufacturers must decide what products to produce.
2. Portfolio managers must decide what investments to purchase while maintaining a portfolio consistent with investors' risk–return preference.
3. Oil company managers must determine, with the help of geologists, where to drill.
4. Corporate managers must choose from among alternative investment projects under conditions of uncertainty.

In this final chapter of the book, we will discuss a variety of statistical methods, collectively referred to as *decision theory*, for dealing with such decision-making problems. First, we present the four key elements of making a decision on the basis of extreme values, expected monetary value, and utility analysis. Then we explore Bayes' strategies and decision trees of expected monetary values in terms of statistical concepts and methodology. We also propose a mean and variance trade-off analysis to replace the expected utility analysis in business decision making. Finally, the mean and variance method is applied in the context of a capital

budgeting decision. Appendix 1 discusses how the spreadsheet can be used to do decision-tree analysis. Appendix 2 presents the graphical derivation of the capital market line; Appendix 3 discusses present value and the net present value (NPV) decision rule; and the derivation of standard deviation for NPV is presented in Appendix 4.

21.2 Four Key Elements of a Decision

Four elements are needed to analyze a decision-making problem: actions, states of nature, outcomes, and probabilities.

1. The choices available to the decision maker are called *actions* (or sometimes *alternatives*). For instance, in a person's decision whether to carry an umbrella, the possible actions are to carry the umbrella and not to carry the umbrella. Although our approach assumes that the decision maker can specify a finite number of mutually exclusive and exhaustive actions, it is also possible to analyze problems with an infinite number of outcomes.
2. The uncertain elements in a problem are referred to as *states of nature*. The states of nature are simply *events*. Like actions, states of nature can be either finite or infinite. The states of nature (events) in our umbrella decision are rain and no rain.
3. An *outcome* is a consequence for each combination of an action and an event (state of nature). The possible outcomes in our umbrella decision are stay dry, be burdened unnecessarily, get wet, and be dry and free. The reward or penalty attached to each outcome is termed the *payoff*, which in business decisions is usually expressed in monetary terms. The relationship among the actions, events, and outcomes involved in a decision process can be presented in a decision tree (see Fig. 21.1).[1]
4. *Probability*, which we discussed in Chap. 5, is the chance that an event will occur. The probability of each event may be set by referring to historical data, expert opinion, or any other factor (including personal judgment) the decision maker wishes to use. These probabilities are initial probabilities, so they are called *prior probabilities*. (Revised probabilities based on the prior probabilities are called *posterior probabilities*. They will be discussed in detail in Sect. 21.5) In our umbrella example, the probability is the chance of rain as assigned in accordance with the weather forecast.

Armed with all the foregoing information, the decision maker can decide whether to carry an umbrella.

In the next section, we first discuss alternative business decision rules without probabilities and then introduce the probability variable into the decision process.

[1] Decision trees were introduced in Sect. 5.3 of Chap. 5. They will be discussed in terms of expected monetary values in Sect. 21.6.

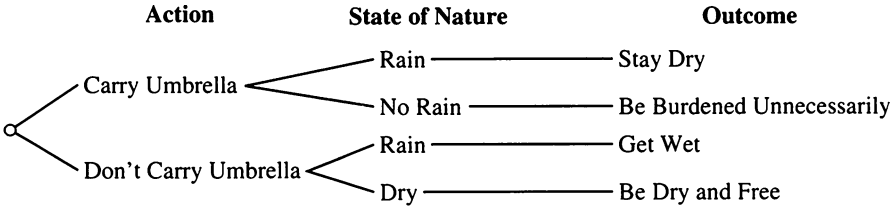


Fig. 21.1 Decision tree for the umbrella decision

Table 21.1 Data for Example 21.1

	State of nature			
Product	*Recession*	*Flat*	*Boom*	Minimum payoff
1	−$10	−$5	$20	−$10
2	−$15	0	$15	−$15
3	−$7	−$1	$25	−$7
4	−$3	$2	$17	−$3

21.3 Decisions Based on Extreme Values

Most of the decision rules discussed in this chapter require the specification of probabilities for the states of nature. However, two strategies do not. The maximin criterion maximizes the minimum payoff; the minimax regret criterion minimizes the difference between the optimal payoff and the actual outcome.

21.3.1 Maximin Criterion

Using this criterion, we first consider the worst possible outcome for each action and then choose the highest of the minimum payoffs (hence "maximin"). The *worst outcome* is simply the smallest payoff that could conceivably result whatever state of nature happens. Thus, the *maximin criterion* is a decision rule for born pessimists. Pessimists tend to assume nature is against them no matter what activity they choose. (Some even go so far as to be convinced that they can guarantee sunny weather by carrying an umbrella!)

Example 21.1 Applying the Maximin Criterion to Different Possible Economic Conditions. Suppose a firm can produce any one of four products, 1, 2, 3, and 4. The four products, then, are the possible actions. The state of nature (which in this case is the state of the economy) has three possibilities: recession, flat, and boom. The payoff, or the outcome, is whatever profits result. This information is presented in Table 21.1.

In this example, the minimum payoff is −$10 for product 1, −$15 for product 2, −$7 for product 3, and −$3 for product 4. Product 4 has the lowest negative payoff in terms of the maximin criterion. Therefore, if it subscribes to the maximin criterion, this pessimistic firm will choose to produce product 4.

Table 21.2 Data for Example 21.2

	State of nature		
	Rain	*No rain*	Minimum payoff
Indoors	$100,000	$100,000	$100,000
Outdoors	0	$100,000,000	0

Table 21.3 Choice of two stocks in terms of rates of return by the maximin criterion

	State of nature			
Stock	*Recession* (%)	*Flat* (%)	*Boom* (%)	Minimum payoff (%)
A (growth type)	−5	8	15	−5
B (income type)	2	4	7	2

Example 21.2 Applying the Maximin Criterion in Rock Music Promotion. Suppose a rock music promoter has the option of booking a major group at the local indoor stadium, which has a seating capacity of 17,000, or at a 100,000-seat outdoor stadium. Obviously, rain is a major concern for the promoter. The possible actions are holding the concert indoors and holding it outdoors; the states of nature are rain and no rain. The payoffs are the profits. These factors are summarized in Table 21.2. By the maximin criterion, the concert should be held indoors because the minimum profits for that strategy are $100,000, whereas they are 0 for holding it outdoors.

The main problem with the maximin criterion is that it does not take into account the probabilities of the states of nature. In the concert example, suppose the concert is to be held in Arizona, where the likelihood of rainfall is quite low. Because the probability of rain is small, it may be better to hold the concert outside.

Another problem with this method is that it does not take other alternatives into consideration. For example, assume that an investor must choose between two stocks and that the states of nature are a recession, a flat economy, and a boom (see Table 21.3). Rates of return are the payoffs. The maximin criterion would dictate choosing stock B, because its minimum payoff is 2 %. However, although stock A has lower rates of return (−5 %) for the recession, its rates of return are much higher when the economy is flat or booming. Of course, the chances of recession, flat economy, and boom are not necessarily equal. Hence, the decision method described in Table 21.3 is relatively restrictive. Later in this chapter, we will explicitly take into account the probability of occurrence of different states of nature.

21.3.2 Minimax Regret Criterion

In the *minimax regret criterion*, the best action is the one that minimizes the maximum regrets for each decision. When the decision maker aims to maximize the benefit, the regret equals the difference between the optimal payoff and the actual payoff. In Example 21.1, the best outcome in the recession is a loss of $3 for product 4, so the regret for product 1 if a recession occurs is $-\$3 - (-\$10) = \$7$, and for

Table 21.4 Regret table

	State of nature			
Product	*Recession*	*Flat*	*Boom*	Maximum regret
1	7	7	5	7
2	12	2	10	12
3	4	3	0	4
4	0	0	8	8

product 2 the regret is $-\$3 - (-\$15) = \$12$. In the flat state of nature, the regret for product 1 is $\$2 - (-\$5) = \$7$, and in the boom state of nature, the regret for product 1 is $\$25 - \$20 = \$5$. The regrets for all four products under the three economic conditions are presented in Table 21.4.

The best product under this criterion is the one that minimizes the maximum regret (hence "minimax"). Thus, product 3 is the best choice because its regret, 4, is the smallest.

Example 21.3 Applying the Minimax Regret Criterion. The following table gives the regrets for the concert example:

	State of nature		
	Rain	No rain	Maximum regret
Indoor	0	900,000	900,000
Outdoor	100,000	0	100,000

By this criterion the best choice is the outdoor concert, because it has the minimum regret ($100,000). Different methods, it seems, can lead to different answers.

The methods presented in this chapter are only aids in decision making. The decision maker must rely partly on personal judgment when making decisions. In our concert example, for instance, the promoter must also take into consideration such factors as the probability of rain, the expected attendance, and ticket prices.

In the following sections, we will discuss adding probability information to statistical decision theory.

21.4 Expected Monetary Value and Utility Analysis

In this section, we discuss both the expected monetary value criterion and utility analysis through probability information. We also examine the application of these techniques in decision making.

Table 21.5 State of economy and payoff

	State of nature			
	Recession	*Flat*	*Boom*	EMV
Probability of occurring	.2	.5	.3	
Product 1	−$20	$0	$15	$.5
Product 2	−$5	−$2	$5	−$.5
Product 3	−$10	$1	$25	$6.0
Product 4	−$1	$5	$0	$2.3

21.4.1 The Expected Monetary Value Criterion

The monetary value at a particular state of nature is calculated by multiplying the probability of the action by the payoff of that particular action. If a decision maker has H possible actions, $A_1, A_2, \ldots, A_H$, and is faced with M states of nature, then the *expected monetary value* associated with the ith action, EMV(A_i), can be obtained by summing the monetary value over all states of nature:

$$\text{EMV}(A_i) = \sum_{j=1}^{M} P_j M_{ij} (i = 1,\ 2, \ldots,\ H) \tag{21.1}$$

where

P_j = probability associated with state of nature j with $\sum_{j=1}^{M} P_j = 1$
M_{ij} = payoff corresponding to the ith action and the jth state of nature

For example, suppose economists estimate the probability of a recession to be .2, that of a flat economy to be .5, and that of a boom to be .3, as shown in Table 21.5. The payoff (profit) and EMV related to each product listed in the last column of Table 21.5 can be calculated as

$$\begin{aligned}
\text{EMV}_1 &= (.2)(-20) + (.5)(0) + (.3)(15) = \$.5 \\
\text{EMV}_2 &= (.2)(-5) + (.5)(-2) + (.3)(5) = -\$.5 \\
\text{EMV}_3 &= (.2)(-10) + (.5)(1) + (.3)(25) = \$6.0 \\
\text{EMV}_4 &= (.2)(-1) + (.5)(5) + (.3)(0) = \$2.3
\end{aligned}$$

The best alternative is the one that maximizes the EMV. In this example, the product with the highest EMV is product 3, with an EMV of $6.0.

Example 21.4 Applying the EMV Criterion to Pricing a New Product. A marketing manager who is responsible for pricing a new product must decide which of the following three alternative pricing strategies to use:

A_1	(Skim-pricing strategy)	$15.50/unit
A_2	(Intermediate price)	$12.00/unit
A_3	(Penetration strategy)	$.50/unit

Table 21.6 State of nature and payoff (thousands of dollars)

	State of nature		
Alternative	*Light demand, S_1*	*Moderate demand, S_2*	*Heavy demand, S_3*
A_1	100	60	−60
A_2	60	110	−30
A_3	−50	0	90
Probability of occurring	.70	.20	.10

The payoff results given in Table 21.6 are total net income associated with each state of nature. In addition, the probability that each state of nature will occur is given at the bottom of the table.

The payoff (net income) is calculated as follows:

$$\text{EMV}(A_1) = (.7)(100) + (.2)(60) + (.1)(-60) = \$76$$
$$\text{EMV}(A_2) = (.7)(60) + (.2)(110) + (.1)(-30) = \$61$$
$$\text{EMV}(A_3) = (.7)(-50) + (.2)(0) + (.1)(90) = -\$26$$

Alternative A_1 offers the highest EMV. It is the best choice if the decision maker's goal is to maximize expected return. However, if the decision maker wants to minimize potential loss, then alternative A_2, with a maximum loss of \$30, is the best choice. In Sect. 21.6, after we discuss Bayes' strategies, this example will be extended to allow sample information.

Example 21.5 Applying the EMV Criterion to Selecting a Stock. A portfolio manager predicts the following probabilities (P_j) for the rates of return on four different stocks associated with three different economic conditions. The rates of return for the jth stock ($i = 1, 2, 3, 4$) are the payoffs.

Using Eq. 21.1, we can calculate the EMV for each stock; all are listed in the last column of Table 21.7. By the EMV criterion, the best choice is stock 4, with a rate of return of

$$(.3)(.20) + (.5)(.03) + (.2)(-.25) = 2.5\%$$

The problem with the EMV criterion is that it does not take the element of risk into consideration. The following example illustrates this. Assume that a coin is flipped. In game 1, a head pays \$2 and a tail pays \$0. In game 2, a head pays \$1 million and a tail pays −\$750,000:

	State of nature		
Game	Heads	Tails	EMV
1	\$2	0	\$1
2	\$1 million	−\$750,000	\$125,000

In the first game, the EMV is (\$2)(.5) + (0)(.5) = \$1; in the second game, it is (\$1,000,000)(.5) + (−\$750,000)(.5) = \$125,000. By the EMV criterion, the best choice is game 2. However, a very high degree of risk is associated with game 2:

Table 21.7 States of market and payoffs

	State of nature			
	Boom	*Flat*	*Recession*	EMV
P_j	.3	.5	.2	
R_1	10 %	0 %	−5 %	2 %
R_2	15 %	−2 %	−10 %	1.5 %
R_3	7 %	2 %	−5 %	2.1 %
R_4	20 %	3 %	−25 %	2.5 %

Table 21.8 Utility function

Scoops	Utility	Marginal utility
1	10	10
2	14	4
3	17	3
4	19	2
5	20	1
6	20.5	.5

$750,000 will be lost if a tail results. In contrast, there is no possibility of loss in game 1. Anyone who chooses game 2 must be prepared to accept a negative expected payoff as the price for the chance of earning a large payoff. In doing so, he is expressing a preference for risk. By contrast, anyone who chooses game 1 accepts a lower expected payoff in order to eliminate the chance of experiencing a large loss—and thus expresses an aversion to risk. The next section shows how we can employ utility analysis to take individual attitude toward risk into account. In Sect. 21.7, we will use mean and variance trade-off analysis to take this kind of investigation further.

21.4.2 Utility Analysis

In all the decisions we have looked at, the decision criterion of choice was the maximization of expected monetary value. That is, an individual or corporation believes that the action offering the highest expected monetary value is the preferred course. However, this kind of decision rule does not allow for risk. For example, investors who, in spreading their investments over a portfolio of stocks, accept a lower expected return in order to reduce the chance of a large loss are expressing an aversion to risk. Hence, the investors' or the managers' attitudes toward risk are important in their decision making.

Utility analysis gives us information on the decision makers' attitude toward risk. Utility measures the satisfaction a consumer or decision maker derives from consumption or the income associated with investment. For example, Table 21.8 gives the utility function for a consumer who consumes ice cream. In this case, the utility function relates the scoops of ice cream consumed to the utility generated from this consumption.

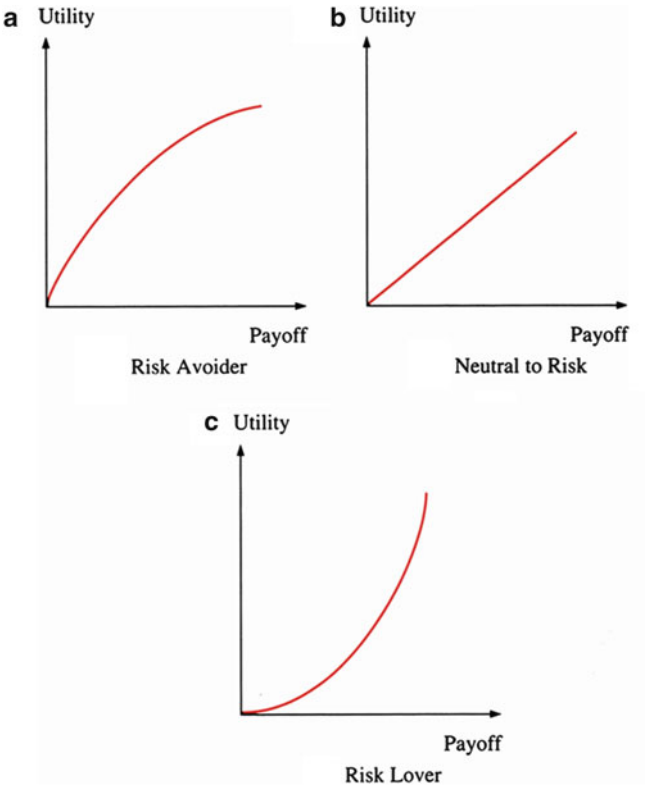


Fig. 21.2 Various types of utility functions

It does not matter what units are used to measure utility. The total utility for the first scoop of ice cream for this consumer could be 100, 140 for the second, and so on. The *total utility* increases as the scoops increase; however, this total utility increases at a decreasing rate. This implies that *marginal utility* decreases as the number of scoops of ice cream increases. That is, as a person eats more and more ice cream, the extra satisfaction received from each additional scoop decreases. In Table 21.8 the marginal utility of any row is equal to the difference of the utility of that row and the utility of the previous row, for example, $4 = 14-10$.

Utility functions can also be used to do utility analysis in business decisions. In this case the *utility function*, a curve relating utility to payoff, can be used to determine whether an investor is risk-averse, risk-neutral, or a risk lover. Various types of utility functions are described in Fig. 21.2. Here the payoff presented on the horizontal axis can originate from either positive or negative value. A *risk-averse* investor has a utility function wherein utility increases at a decreasing rate as payoff increases. In other words, a risk avoider prefers a small but certain monetary gain to a gamble that has a higher expected monetary value but may involve a large but unlikely gain or a large but unlikely loss. A *risk-neutral* investor has a utility function wherein utility increases at a constant rate. For an individual neutral to

Table 21.9 Alternative investment payoffs

Investment opportunities	Payoffs	Probability	Utility
Risky	−$100	. 5	0
	$200	.5	1
Risk-free	$50	1.0	?

risk, every increase of, say, $100 has an associated constant increase in utility. This type of individual uses the criterion of maximizing expected monetary value in decision making, because doing so maximizes expected utility. A *risk lover*'s utility function has utility increasing at an increasing rate. This type of person willingly accepts gambles having a smaller expected monetary value than an alternative payoff that is a "sure thing."

We will use the following example to analyze how the utility function operates within the decision-making process and how it affects the decision.

Suppose an investor faces investment opportunities with the payoffs −$100, $200, and $50, as indicated in Table 21.9. We are interested in the investor's utility level for each situation. The different utility levels the investor can reach lead to different decisions. For simplicity, we attach a utility of 0 to the payoff of −$100 and a utility of 1 to the $200 payoff, leaving us with the utility for the $50 payoff. In order to link the utility for the $50 payoff with the information on the decision maker's preference for risk, we then ask, "Would the investor prefer to receive $50 with certainty or to gamble, possibly gaining $200 but just as likely to lose $100?"

Drawing the information given in Table 21.9, we can calculate the expected utility of the payoff as

$$.5U(-\$100) + .5U(\$200) = .5(0) + .5(1) = .5$$

Case I: Risk-averse

$$U(\$50) > .5$$

Case II: Risk-neutral

$$U(\$50) = .5$$

Case III: Risk lover

$$U(\$50) < .5$$

Thus, we know that if the utility for the $50 payoff is greater than .5, the investor is risk-averse. By similar reasoning, if the utility for the $50 payoff is less than (equal to) .5, the investor is a risk lover (is risk-neutral). We graph these three alternative utility curves in Fig. 21.3, which is a more detailed version of Fig. 21.2.

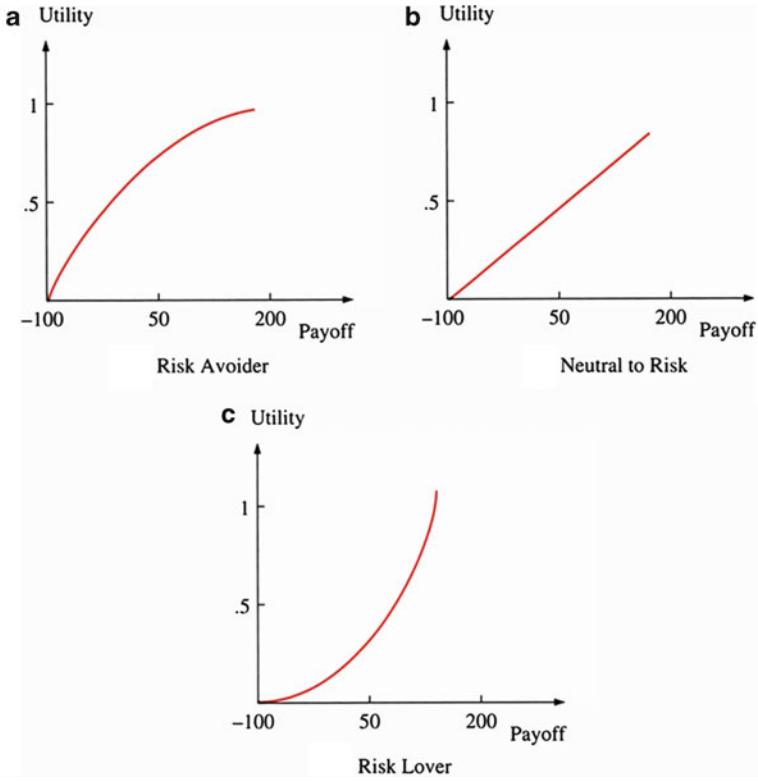


Fig. 21.3 Alternative utility curves

The expected monetary value (EMV) for risky investment opportunities with two possible uncertain investment payoffs as indicated in Table 21.9 is

$$(-\$100)(.5) + (\$200)(.5) = \$50$$

If we use the EMV approach to analyze this problem, we implicitly assume that the investor is risk-neutral. If we employ utility analysis, we consider not only the risk-neutral case but also the cases of both the risk avoider and the risk lover. In short, the EMV approach uses expected *objective* dollar utility as the decision criterion, and utility analysis uses expected *subjective* utility.

The process we went through in this example tells us several things:

1. The utility function affects the decision-making process.
2. If the utility of the expected payoff is greater than, equal to, or less than the expected utility of the payoff, the decision maker will be risk-averse, risk-neutral, or risk-loving, respectively.[2]

[2] In our case, the expected utility of the payoff is 0.5. The utility of the expected payoff for a risk-averse investor is larger than .5; the utility of the expected payoff for a risk-neutral investor is .5; and the utility of the expected payoff for a risk lover is smaller than .5.

3. If the utility function is strictly concave, linear, or convex, the decision maker will be risk-averse, risk-neutral, or risk-loving, respectively.
4. If marginal utility is decreasing, constant, or increasing, the decision maker will be risk-averse, risk-neutral, or risk-loving, respectively.

Let's look at another example of the use of utility analysis in decision making.

Example 21.6 Different Attitudes Toward Risk. Assume that an investor's utility function can be defined by $U(x) = \sqrt{x}$, where $x =$ return (payoff). Now, he faces the choice of $x = 14.5$ with certainty or $x = 25$ with probability .5 and $x = 4$ with probability .5. We want to determine whether this is a risk-averse person.

From the foregoing example, we know that if the investor's utility of the expected payoff is greater than the expected utility of the payoff, if the utility function is strictly concave, or if the marginal utility is diminishing, then the investor is said to be risk-averse. In fact, the last two criteria are the same. To determine whether this individual is risk-averse, we simply take the ordinary derivatives of $U(x)$ with respect to x two times; then we can mathematically show that the individual is risk-averse if x is larger than zero.[3] Such an investor will take the certain payoff of $x = 14.5$.

Another approach to determining an individual's risk preference is to calculate the expected utility of the payoff and the utility of the expected payoff as indicated in Table 21.10. In the table, $U(x)$, the utility given x, is calculated by substituting the values for x of 4 and 25, respectively, into the utility function $U(x) = \sqrt{x}$. $E[U(x)]$, the expected utility given x, is equal to $pU(x)$. Finally, $E(x)$, the expected value of x, and $U\,[E(x)]$, the utility from the expected value of x, are calculated as follows:

$$E(x) = (.5)(4) + (.5)(25) = 14.5$$
$$U[(E(x))] = \sqrt{14.5} = 3.808$$

From this example, we can see that the utility of the expected payoff, 3.808, is greater than the expected utility of the payoff, 3.5 (1 + 2.5). Again, we can see that the investor is risk-averse; the utility received from the expected value of x is greater than the expected value of the utility of x. What this means is that the investor will get greater utility from receiving the expected value of x with certainty than he'd get if he took the gamble.

3

$$\frac{dU(x)}{dx} = \frac{d(x)^{1/2}}{dx} = \tfrac{1}{2}x^{-1/2}$$
$$\frac{d^2U(x)}{dx^2} = \frac{d(\frac{1}{2}x^{-1/2})}{dx} = -\tfrac{1}{4}x^{-3/2}$$

If $x > 0$, then we know the second derivative of the utility function is negative. This condition represents the curve as concave, as shown in Fig. 21.3a.

Table 21.10 Expected utility of the payoff and utility of the excepted payoff

x	$U(x) = \sqrt{x}$	P	$E[U(x)]$	$E(x)$	$U[E(x)]$
4	2	.5	1	2	
25	5	.5	2.5	12.5	
–	–	1.0	3.5	14.5	3.808

An investor is risk-averse if he prefers a safe investment over a risky investment with a higher expected value. For example, if an investor prefers to invest in risk-free Treasury bills rather than investing in a stock with a higher expected value, he is considered risk-averse.

Having determined the appropriate utilities, we need only solve the decision-making problem by finding that course of action with the highest *expected utility*. Employing utility analysis concepts, we can modify the expected monetary value criterion defined in Eq. 21.1 to be the expected utility criterion:

$$E[U(A_i)] = \sum_{j=1}^{m} P_j U_{ij} (i = 1, 2, \ldots, n) \tag{21.2}$$

where

$E[U(A_i)]$ = expected utility of action i
P_j = probability associated with state of nature j
U_{ij} = utility corresponding to the ith action and the jth state of nature

In Eq. 21.2, we also assume that $\sum_{j=1}^{m} P_j = 1$. If the decision maker is risk-neutral (indifferent to risk), the expected utility criterion and the expected monetary value criterion are equivalent.

21.5 Bayes' Strategies

We studied Bayes' theorem in Chap. 5. This theorem enables us to work out the probability for one event that is conditional on another event. Bayesian analysis can be used in the decision-making process. The difference between Bayes analysis, the maximin criterion, and the minimax regret criterion is that in Bayes analysis, probabilities of the states of nature must be specified.

Recall Bayes' theorem:

$$P(E_2|E_1) = \frac{P(E_1|E_2)P(E_2)}{P(E_1)} \tag{21.3}$$

where $P(E_2|E_1)$ and $P(E_1|E_2)$ are, respectively, conditional probabilities of event 2 (E_2) given event 1 (E_1) and of E_1 given E_2. $P(E_1)$ and $P(E_2)$ are unconditional probabilities of E_1 and E_2, respectively. In terms of Bayesian statistic, $P(E_2)$ is the

initial or prior probability of E_2 and is modified to the posterior probability, $p(E_2|E_1)$, given the sample information that event E_1 has occurred. We can incorporate different states of nature into Eq. 21.3 to obtain the generalized Bayes model defined in Eq. 21.4:

$$P(S_i|I) = \frac{P(I \cap S_i)}{P(I)} = \frac{P(I|S_i)P(S_i)}{\sum_{i=1}^{m} P(I|S_i)\, P(S_i)} \tag{21.4}$$

where $P(S_i/I)$ is the probability of state of nature S_i given sample information I. $P(S_i)$ is the probability of state of nature S_i *not* incorporating sample information I, and it is called a prior probability of S_i. We also assume that there exist m states of nature.

Example 21.7 Bayesian Approach in Forecasting Interest Rates. Suppose that macroeconomists are hired to predict interest rates. Past results for economic prognostications are presented in the following table.

	Interest rate outcome	
Belief	Up	Down
Strong credit market	.60	.30
Weak credit market	.40	.70

When economists believed that credit markets would be strong, interest rates went up and went down with 60 % and 30 % chances, respectively. Thus,

$$\begin{aligned} P(\text{strong market/up}) &= .60 \\ P(\text{strong market/down}) &= .30 \end{aligned}$$

Now suppose the economists believe that the probability that rates will rise is .7 and that the probability of lower rates is .3.

$$P(\text{up}) = .7 \qquad P(\text{down}) = .3$$

Following Eq. 21.4, we find that in the case of two states of nature, the probability that interest rates will rise, given a strong credit market assessment by economists, is

$$\begin{aligned} &P(\text{up}|\text{strong market}) \\ &= \frac{P(\text{strong market}|\text{up})P(\text{up})}{P(\text{strong market}|\text{up})[P(\text{up})] + P(\text{strong market}|\text{down})[P(\text{down})]} \\ &= \frac{.60(.70)}{.6(.7) + .3(.3)} = \frac{.42}{.51} = .82 \end{aligned}$$

The probability that interest rates will rise, given a weak credit market assessment, is

$$P(\text{up/weak market}) = \frac{.4(.7)}{(.4)(.7) + (.7)(.3)} = .57$$

Corporate executives can use these probabilities to assess future interest rates.

21.6 Decision Trees and Expected Monetary Values

In this section we use the expected monetary value (EMV) criterion in decision-tree form to select the best alternative in business decision making. As a general approach to structuring complex decisions, a decision tree helps direct the user to a solution. It is a graphical tool that describes the types of actions available to the decision maker and the resulting events.

The decision-tree approach to capital budget decision making is used to analyze investment opportunities involving a sequence of investment decisions over time. To best illustrate the use of the decision tree, we will develop a problem involving numerous decisions.

First, we must enumerate some of the basic rules for implementing this method. The decision maker should try to include only important decisions or events. The decision-tree model requires the decision maker to make subjective estimates when assessing probabilities. And it is important to develop the tree in chronological order to ensure the proper sequence of events and decisions.

A decision point, or decision node, is represented by a box. The available alternatives are represented by branches out of this node. A circle represents an *event node*, and branches from this type of node represent possible events.

The expected monetary value (EMV) is calculated for each event node by multiplying probabilities by conditional profits and summing them. The EMV is then placed in the event node and represents the expected value of all branches arising from that node.

A decision tree is shown in Fig. 21.4. The states of nature are high, medium, and low levels of GNP, and their probabilities are .2, .5, and .3, respectively. For product l, the expected value is $(.2 \times 100) + (.5 \times 5) + (.3 \times -30) = 13.5$. The highest EMV (14) is that of product 3.

Each square on the decision tree denotes a decision that must be made. The circles indicate the states of nature. The square represents the decision to produce product 1, 2, or 3. As we have said, the states of nature that can occur are high, medium, and low levels of economic performance. Now let's look at two examples of how decision trees using objective (dollar payoff) utility instead of subjective utility are employed in business decision making.

Example 21.8 A Decision Tree for Testing a Drilling Site. An oil company is trying to decide whether to test for the presence of oil or to drill for oil (see Fig. 21.5). If oil is struck, the revenues are $1 million, with a cost of $100,000 to drill. The firm has to decide whether to test first for the presence of oil. Without testing, the probability of striking oil is .1. Thus, the firm's expected value without testing is 0 ($1 million

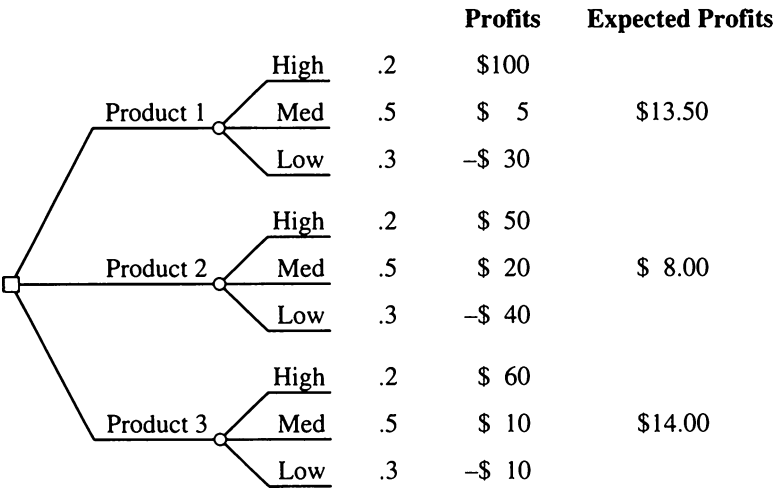


Fig. 21.4 Decision tree for determining expected profit

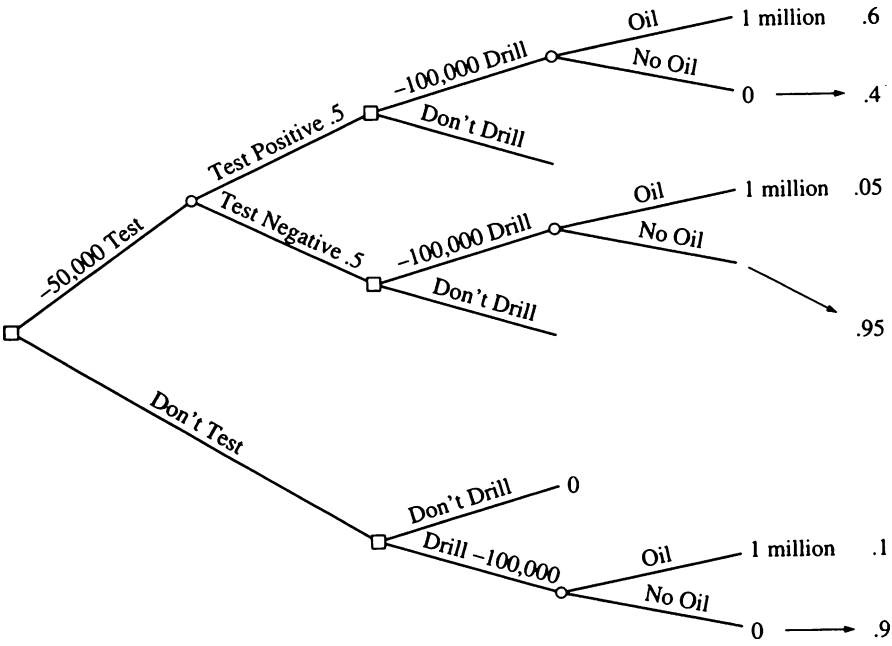


Fig. 21.5 Decision tree for Example 21.8

times .1, less drilling fees of $100,000). The cost of the oil test is $50,000. If the test is positive, there is a .6 probability that oil will be struck; if the test is negative, the probability of striking oil is .05.

The expected gain for a negative test result is .05(1,000,000) − 100,000 − 50,000 = −$100,000. If the test is positive, the expected profit is (1,000,000)(.6) − 100,000 − 50,000 = $450,000. If the firm's test is negative, the firm should not drill; however, if the test is positive, it should drill. This analysis makes it clear that the test should be done. In Appendix 1 we show how the spreadsheet can be used for decision-tree analysis for drilling oil.

Example 21.9 A Decision Tree for Capital Budgeting. A firm currently sells paper and paperboard packaging materials. Company planners predict that, with the advent of plastic shrink-film packaging, their line of products may be obsolete within a decade. They must quickly decide on a short-term plan of action from among four alternatives: (1) do nothing, (2) establish a tie-in with a machinery company that manufactures plastic packaging, (3) acquire such a company, or (4) take on the research and development of plastic packaging. These four alternatives are the first four branches arising out of the event node in Fig. 21.6. If the company planners do nothing, the firm's short-term profits will be about the same as in the previous year. If they decide to establish a tie-in with another firm, they foresee one of two events occurring; there is a 90 % chance of successful introduction of their new plastics line and a 10 % possibility of failure. If they decide on acquisition, they foresee a 10 % chance of problems with antitrust laws, a 30 % possibility of an unsuccessful introduction of the plastics line, and a 60 % chance of success. If they decide to manufacture a whole plastics line on their own, they foresee many more problems. They anticipate a 10 % chance of having trouble developing their own machines, a 10 % chance of having problems with suppliers in developing a total packaging system for their customers, a 30 % chance that customers will not purchase their systems, and a 50 % chance of success in the development and introduction of the plastics line.

Conditional profit is the amount of profit the firm can expect to make by adopting each of the preceding sets of alternatives and consequent events.

In Fig. 21.6 the expected monetary values are shown in the event nodes. The firm's financial planner can use EMV to decide which action to take, selecting the decision node with the highest EMV (in this case, establishing a tie-in, which has an EMV of 76.5). The slash marks indicate elimination of nonoptimal decision branches from consideration. If the probabilities associated with events change, then the EMVs associated with the alternatives may change—and with them the selection of the optimal alternative.

For Example 21.9, we greatly simplified the number of possible alternatives and events. In fact, decision trees are more useful for more complex problems—that is, for problems containing more possibilities or problems in which management must make a sequence of decisions rather than a single decision.

Example 21.10 Utilizing Sample Information to Improve the Determination of Pricing Policy. In Example 21.4 we determined the price of a new product without

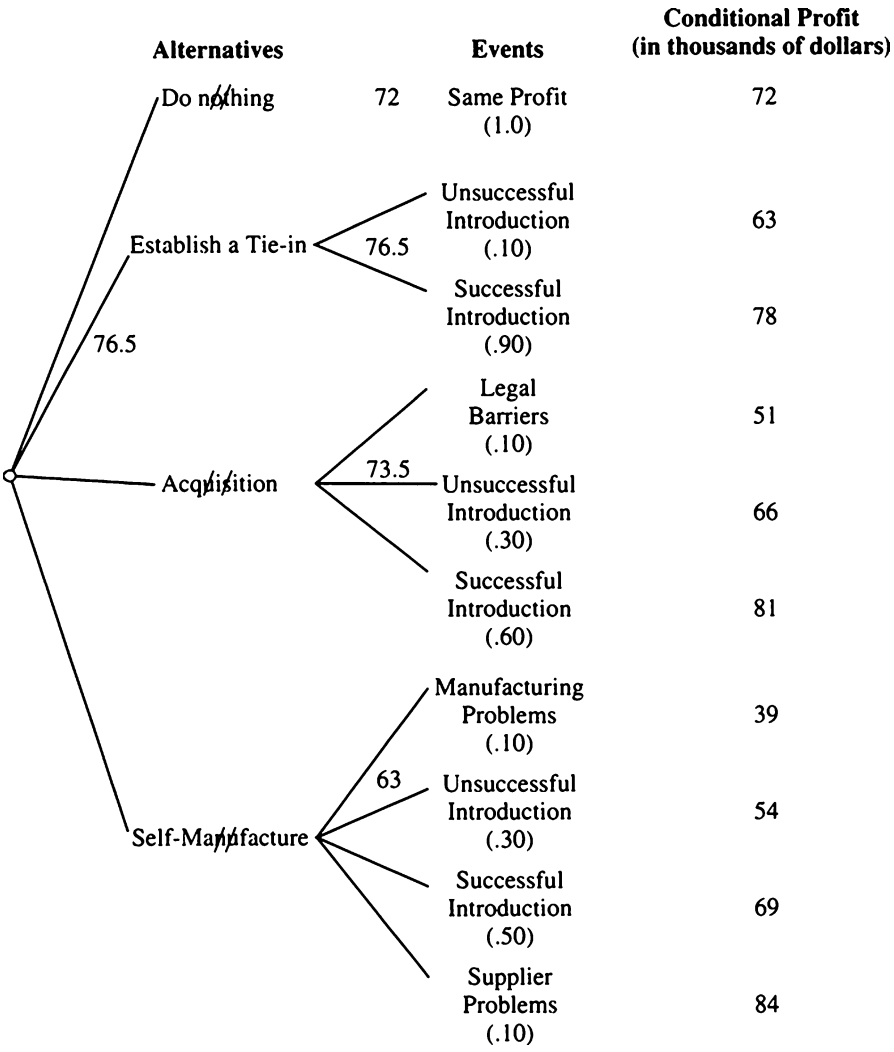


Fig. 21.6 Decision tree for Example 21.9 (*Source:* Cheng F. Lee and Joseph E. Finnerty (1990), Corporate Finance, Theory, Method, and Application (San Diego: Harcourt Brace Jovanovich))

using sample information.[4] Such sample information as test market results, however, can be very helpful. Suppose past product performances can give some indication about the relationship between test market results and product performance nationally. Let

[4] This example is similar to an example given in Gilbert A. Churchill, Jr. (1983), *Market Research: Methodological Foundations*, 3d ed. (Chicago: Dryden) pp. 37–42.

Table 21.11 Conditional probabilities of each test market result, given each state of nature

	State of nature		
Test market result	*Light demand* S_1	*Moderate demand* S_2	*Heavy demand* S_3
Z_1	.5	.2	.2
Z_2	.3	.7	.6
Z_3	.2	.1	.2
	1.0	1.0	1.0

Table 21.12 Revision of prior probabilities in light of possible test market result

(1) j	(2) State of nature S_j	(3) Prior probability $P(S_j)$	(4) Conditional probability $P(Z_k\|S_j)$	(5) = (3) × (4) Joint probability $P(S_j)P(Z_k\|S_j)$	(6) = (5) ÷ sum of (5) Posterior probability $P(S_j\|Z_k)$
Z_1	S_1	.7	.5	.35	.854
	S_2	.2	.2	.04	.097
	S_3	.1	.2	*.02*	*.049*
				.41	1.000
Z_2	S_1	.7	.3	.21	.512
	S_2	.2	.7	.14	.342
	S_3	.1	.6	*.06*	*.146*
				.41	1.000
Z_3	S_1	.7	.2	.14	.778
	S_2	.2	.1	.02	.111
	S_3	.1	.2	*.02*	*.111*
				.18	1.000

Z_1 = disappointing or only slightly successful test market
Z_2 = moderately successful market
Z_3 = highly successful market

By using Bayes' theorem (Eq. 21.4) and supposing that past experiences provided the estimate of conditional probabilities given in Table 21.11, we find the revised prior probabilities $P(S_1) = .7$, $P(S_2) = .2$, and $P(S_3) = .1$, as presented in Table 21.12. Conditional probabilities from Table 21.11 are presented in column (4) of Table 21.12. Using Eq. 21.4, we calculate the posterior probabilities $P(S_j/Z_k)$ presented in column (6) of Table 21.12. Using the information on states of nature and payoffs listed in Table 21.6 and the posterior probability information listed in Table 21.12, we find the expected value of each alternative, given each research outcome (see Table 21.13).

The probability of obtaining each test market result—that is, the probability of each Z_k—is given as

$$P(Z_k) = \sum_{j=1}^{n} P(S_j)P(Z_k|S_j)$$

and for k—1, for example, the probability is

Table 21.13 Expected value of each alternative, given each research outcome

Z_1, disappointing or only slightly successful test market
$\text{EV}(A_1) = (100)(.854) + (60)(.097) + (-60)(.049) = 88.28$
$\text{EV}(A_2) = (60)(.854) + (110)(.097) + (-30)(.049) = 60.44$
$\text{EV}(A_3) = (-50)(.854) + (0)(.097) + (90)(.049) = -38.29$
Z_2, moderately successful test market
$\text{EV}(A_1) = (100)(.512) + (60)(.342) + (-60)(.146) = 62.96$
$\text{EV}(A_2) = (60)(.512) + (110)(.342) + (-30)(.146) = 63.96$
$\text{EV}(A_3) = (-50)(.512) + (0)(.342) + (90)(.146) = -12.46$
Z_3, highly successful test market
$\text{EV}(A_1) = (100)(.778) + (60)(.111) + (-60)(.111) = 77.80$
$\text{EV}(A_2) = (60)(.778) + (110)(.111) + (-30)(.111) = 55.56$
$\text{EV}(A_3) = (-50)(.778) + (0)(.111) + (90)(.111) = -28.91$

$$\begin{aligned} P(Z_1) &= P(S_1)P(Z_1|S_1) + P(S_2)P(Z_1|S_2) + P(S_3)P(Z_1|S_3) \\ &= (.7)(.5) + (.2)(.2) + (.1)(.2) \\ &= .41 \end{aligned}$$

The probability is given as the sum of the elements in column (5) of Table 21.12. Table 21.12 thus indicates that the probabilities associated with these test markets are $P(Z_1) = .41$, $P(Z_2) = .41$,and $P(Z_3) = .18$. (The probabilities sum to 1, as they should, because one of the three test market outcomes must result.) The expected value of the test-marketing procedure is found by weighting each expected value of the optimal action in Table 21.13, given each research result, by the probability of receiving that expected value. The expected value of the proposed research is thus found to be

$$\begin{aligned} \text{EV(research)} &= (88.28)(.41) + (63.96)(.41) + (77.80)(.18) \\ &= 76.42 \end{aligned}$$

This value is \$0.42 over the expected value of the optimal action without research, which, as indicated in Example 21.4, is \$76. Hence, the market research should be undertaken if the research cost is less than \$0.42.

21.7 Mean and Variance Trade-Off Analysis

21.7.1 The Mean–Variance Rule and the Dominance Principle

The expected utility rule we discussed in Sect. 21.4 is theoretically the best criterion available, but sometimes it is very hard to implement. We frequently do not know the investor's utility function, and furthermore, the decision maker, as in the case of a manager, must act on behalf of many stockholders with different utility functions.

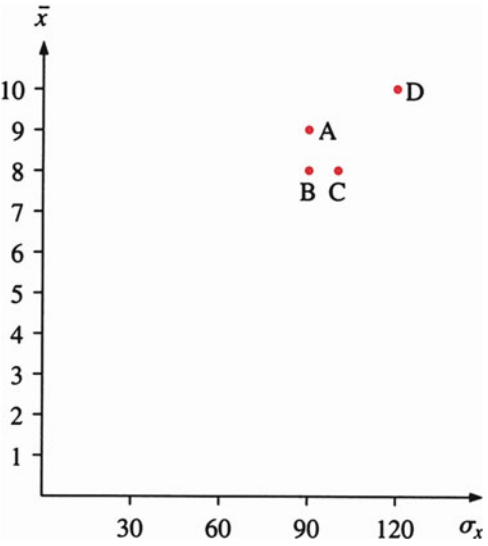


Fig. 21.7 Trade-off between mean and standard deviation for investment projects

Hence, the expected utility rule is often replaced by a more practical mean–variance decision criterion that assumes that the decision maker has a risk-aversion utility function.

According to the mean–variance rule, the expected return (mean) measures an investment's profitability, whereas the variance (or standard deviation) of returns measures its risk. Consider the following four alternative projects, with the means $\bar{x}$ and standard deviations σ_x specified.

Investment project	$\bar{x}$	σ_x
A	\$9	\$90
B	8	90
C	8	100
D	10	120

To discuss the implications of the trade-off between risk and return and of the dominance principle, we plot this set of data in Fig. 21.7. A pairwise comparison of the investment projects shows that project A dominates projects B and C; it has the highest return, and its risk is equal to that of project B and lower than that of project C. However, there is no clear-cut decision between projects A and D, projects B and D, or projects C and D. Here the investor needs to consider the trade-off between profit and risk in terms of his or her attitude toward risk.

In the analysis of stock investments, average rates of return and their variance (or standard deviation) are used to represent investments' profitability and risk, respectively. The variance of rates of return can be decomposed into two components by the market model[5] defined in Eq. 21.5.

[5] We discussed the market model in Chap. 14 and elsewhere.

$$R_{i,t} = \alpha_i + \beta_i R_{m,t} + e_{i,t} \tag{21.5}$$

where $R_{i,t}$ and $R_{m,t}$ are rates of return for the ith security (portfolio) and market rates of return, respectively.

Following Eq. 13.19 of Chap. 13,

$$\sigma_i^2 = \beta_i^2 \sigma_m^2 + \sigma_{ei}^2 \tag{21.6}$$

where

σ_i^2 = variance of $R_{i,t}$

σ_m^2 = variance of market rates of return

σ_{ei}^2 = residual variance of rates of return for the ith security

In investment analysis, we define σ_i^2, $\beta_i^2 \sigma_m^2$, and σ_{ei}^2 as total risk, systematic risk, and unsystematic risk, respectively.

Systematic risk is the part of total risk that results from the basic variability of stock prices. It accounts for the tendency of stock prices to move together with the general market. The other portion of total risk is unsystematic risk, which is the result of variations peculiar to the firm or industry—for example, a labor strike or resource shortage.

Systematic risk, also referred to as *market risk*, reflects the fluctuations and changes in general market conditions. Some stocks and portfolios are very sensitive to movements in the market; others exhibit more independence and stability. A measure of a stock's or a portfolio's relative sensitivity to the market, assigned on the basis of its past record, is designated by the upper-case Greek letter beta (β).

Example 21.11 Market Model and Risk Decomposition for JNJ. The annual rate of return for JNJ and the market rate of return for 1990–2009 are used to estimate the market model in accordance with Eq. 21.5 and in terms of MINITAB. The results are shown in Fig. 21.8. From Fig. 21.8, we find that the beta coefficient for the market model is .639. From the analysis of variance data in Fig. 21.8, we obtain the total risk (σ_i^2), systematic risk ($\beta_i^2 \sigma_m^2$), and unsystematic risk (σ_{ei}^2) as follows:

$$\begin{aligned}
\sigma_i^2 &= \frac{1.74158}{19} = .09166 \\
\beta_i^2 \sigma_m^2 &= \frac{.17720}{19} = .00933 \\
\sigma_{ei}^2 &= \frac{1.56437}{19} = .08234 \\
\beta_i^2 \sigma_m^2 + \sigma_{ei}^2 &= .00933 + .08234 = .09167 \doteq .09166
\end{aligned}$$

Fig. 21.8 MINITAB output of the market model for JNJ

Data Display

Row	JNJ	S&P
1	0.230108	0.036396
2	0.616842	0.124301
3	-0.551293	0.105162
4	-0.091578	0.085799
5	0.244915	0.019960
6	0.584558	0.176578
7	-0.409758	0.237724
8	0.340804	0.302655
9	0.287688	0.242801
10	0.124207	0.222782
11	0.139719	0.075256
12	-0.431191	-0.163282
13	-0.078010	-0.167680
14	-0.021172	-0.028885
15	0.248595	0.171379
16	-0.032496	0.067731
17	0.122480	0.085510
18	0.034602	0.127230
19	-0.076435	-0.174081
20	0.108473	-0.222935

Regression Analysis: JNJ versus S&P

The regression equation is
JNJ = 0.0273 + 0.639 S&P

Predictor	Coef	SE Coef	T	P
Constant	0.02726	0.07227	0.38	0.710
S&P	0.6386	0.4472	1.43	0.170

S = 0.294804 R-Sq = 10.2% R-Sq(adj) = 5.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.17720	0.17720	2.04	0.170
Residual Error	18	1.56437	0.08691		
Total	19	1.74158			

Durbin-Watson statistic = 2.51280

21.7.2 The Capital Market Line

From the rates of return and alternative measurements of risk, we can derive either the trade-off between expected return and total risk or the trade-off between expected return and systematic risk. In these cases, the utility function used to measure a decision maker's attitude toward risk is the von Neumann and Morgenstern (VNM) type. The use of the term *utility* in the VNM type of utility function differs from its use by traditional economists. The VNM type of utility function is applied in situations where money payoffs are inappropriate as a measuring device. In traditional economics, utility reflects the inherent satisfaction delivered by a commodity and is measured in terms of psychic gains and losses. Von Neumann and Morgenstern, on the other hand, conceived of utility as a measure of value that provides a basis for making choices in the assessment of situations involving risk. This approach integrates the EMV and the utility analyses discussed in Sect. 21.4.[6]

The *capital market line* used to describe the trade-off between expected return and total risk is[7]

$$E(R_i) = R_f + \left[E(R_m) - R_f\right] \frac{\sigma_i}{\sigma_m} \tag{21.7}$$

where

R_f = risk-free rate
$E(R_m)$ = expected return on the market portfolio
$E(R_P)$ = expected return on the ith portfolio
σ_i, σ_m = standard deviations of the portfolio and the market, respectively

The capital market line (CML) defined in Eq. 21.7 implies that the expected rates of return for portfolio i equal the risk-free rate plus total market risk,

$$\frac{\left[(E(R_m) - R_f\right]\sigma_i}{\sigma_m}.$$

The total portfolio risk premium is equal to price per market risk,

$$\frac{E(R_m) - R_f}{\sigma_m},$$

[6] In other words, the utility function can be defined as $U[E(R), \sigma]$, where $E(R)$ and σ represent expected rates of return and the standard deviation of rates of return, respectively. By assuming that the investors are risk avoiders, we have $[\partial\, U/\partial\, E(R)] > 0$ and $(\partial\, U/\partial\, \sigma) < 0$. In other words, investors prefer return and dislike risk.

[7] The graphical derivation of this mode appears in Appendix 2.

Table 21.14 Return and standard deviation for mutual funds

	Mutual fund A (%)	Mutual fund B (%)
Average return, $\bar{R}_i$	20	15
Standard deviation, σ_i	8	5

times total risk associated with portfolio i—that is, σ_i. By using the concept of CML, we can define the *Sharpe investment performance measure* for the *i*th portfolio as

$$\text{SM}_i = \frac{\bar{R}_i - R_f}{\sigma_i} \tag{21.8}$$

Example 21.12 Using the Sharpe Investment Performance Measure to Determine Investment Performance. An investor is considering investing in either mutual fund A or mutual fund B. For past performance, he calculates for both funds the average returns and variances listed in Table 21.14. It is assumed that the T-bill rate is 8 %, which the firm uses as the risk-free rate.

The Sharpe performance measure, then, gives

$$\text{SM}_A = \frac{.20 - .08}{.08} = 1.5$$
$$\text{SM}_B = \frac{.15 - .08}{.05} = 1.4$$

These calculations reveal that mutual fund A will give a slightly better performance and thus is the better alternative of the two investments.

21.7.3 The Capital Asset Pricing Model

The capital market line (Eq. 21.7) is used to describe the trade-off between expected rate of return and total risk. The trade-off between expected rate of return and systematic risk defined in Eq. 21.9 is called the *capital asset pricing model* (CAPM):

$$E(R_i) = R_f + \beta_i\big(E(R_m) - R_f\big) \tag{21.9}$$

where

$E(R_i)$ = expected rate of return for asset i
R_f = risk-free rate
β_i = measure of systematic risk (beta) for asset i
$E(R_m)$ = expected return on the market portfolio

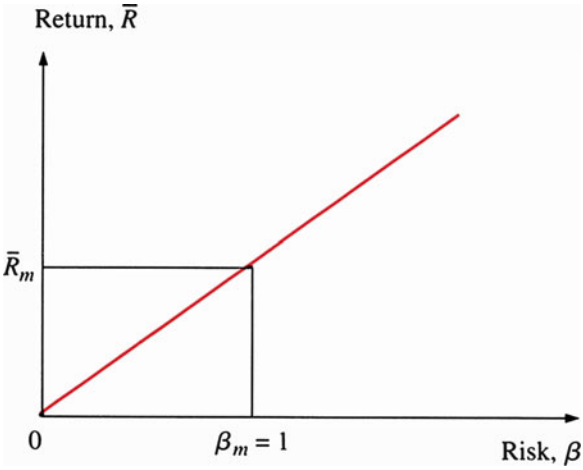


Fig. 21.9 The capital asset pricing model (CAPM)

Equation 21.9 implies that β_i is the systematic risk for determining the price of the individual asset and the portfolio. Figure 21.9 illustrates graphically the relationship between $E(R_i)$ and β_i that is defined in Eq. 21.9. Professor William Sharpe won the Nobel Prize in Economics mainly because he derived this model.

The reason why the CAPM can be regarded as part of decision theory is that it is based on a utility function in terms of expected rates of return and the standard deviation of rates of return. Expected rates of return and the standard deviation of rates of return are essentially based on the monetary value of the investment. In Sect. 21.4, we used expected monetary values and utility analysis to make decisions. Here we treat risk (the standard deviation of rates of return) as an explicit factor, whereas in Sect. 21.4 we treated it as an implicit factor in determining the value of an investment.

With the capital asset pricing model, we must assume that all utility-maximizing investors will attempt to position themselves somewhere along the CML and will attempt to put some portion of their wealth into the market portfolio of risky assets.

The CAPM implies that the market portfolio is the only relevant portfolio of risky assets. Hence, the relevant risk measurement of any individual security is its covariance with the market portfolio—that is, the systematic risk of the security.

The relationship between the capital market line (CML) and the CAPM can be shown by starting with the definition of the beta coefficient:

$$\beta_i = \frac{\text{Cov}(R_i, R_m)}{\text{Var}(R_m)} = \frac{\rho_{i,m}\sigma_i\sigma_m}{\sigma_m^2} = \frac{\rho_{i,m}\sigma_i}{\sigma_m}$$

where

σ_i = standard deviation of the ith security's rate of return
σ_m = standard deviation of the market rate of return
$\rho_{i,m}$ = correlation coefficient of R_i and R_m
If $\rho_{i,m} = 1$, then Eq. 21.9 reduces to

$$E(R_i) = R_f + \frac{\sigma_i}{\sigma_m}\left[\left(E(R_m) - R_f\right)\right] \tag{21.10}$$

If $\rho_{i,m} = 1$, then this implies that the portfolio in question is the efficient portfolio, or, for an individual security, it implies that the returns and risks associated with the asset are similar to those associated with the market as a whole. The implications of this comparison, in turn, are that

1. Equation 21.9 is a generalized case of Eq. 21.10, because Eq. 21.9 includes the correlation coefficient, whereas Eq. 21.10 assumes that the correlation coefficient is equal to 1.
2. The capital asset pricing model (CAPM) instead of the capital market line (CML) should be used to price an individual security or an inefficient portfolio. To use the CML to price an inefficient portfolio would be to price unsystematic risk.
3. The CML prices the risk premium in terms of total risk, and the security market line (SML) prices the risk premium in terms of systematic risk.

In order to apply the CAPM, we need to estimate the beta coefficient, the risk-free rate, and the market risk premium. Estimates of these quantities can be obtained from time-series data as shown in Example 21.11 (see also Chap. 14).

The capital asset pricing model (CAPM) defined in Eq. 21.9 implies that rates of return for the ith security (or portfolio) equal the risk-free rate plus the security's (or portfolio's) risk premium $[E(R_m)$—$R_f\, \beta_i$. This risk premium is equal to systematic risk β_i times the expected market risk premium, $E(R_m$—$R_f)$. By using the concept of CAPM, we can define the *Treynor investment performance measure*[8] for the ith security (or portfolio) as follows:

$$\mathrm{TM}_i = \frac{\bar{R}_i - R_f}{\beta_i} \tag{21.11}$$

If in Eq. 21.12 we also know that the beta coefficients for mutual funds A and B are $\beta_A = 1.8$ and $\beta_B = 1.2$, respectively, then the Treynor investment measures for these two mutual funds are

[8] The derivation and justification of this investment performance measure can be found in J. Treynor (1965), "How to Rate Management of Investment Fund," *Harvard, Business Review* 43, 63–75.

$$\text{TM}_A = \frac{.20 - .08}{1.8} = .0667$$
$$\text{TM}_B = \frac{.15 - .08}{1.2} = .0583$$

Like the Sharpe performance measure, the Treynor measure indicates that mutual fund A is the better alternative of the two investments.

By using the specification of Eq. 21.9, we can define the CAPM version of market model as

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i\left(R_{m,t} - R_{f,t}\right) + e_{i,t} \tag{21.12}$$

where $R_{i,t}$ are the rates of return for ith security (or portfolio) in period t; $R_{m,t}$ are market rates of return in period t; and $R_{f,t}$ is the return on a risk-free asset (such as T-bills rate) in period t. Jensen (1968) has shown that a_i can be used to evaluate the investment performance for either security or portfolio.[9]

Therefore, it is called the Jensen investment performance measure which can be explicitly defined as

$$JM_i = \overline{R}_i - \left[\overline{R_f} + \beta_i\left(\overline{R_m} - \overline{R_f}\right)\right] \tag{21.13}$$

where $\bar{R}_i, \bar{R}_m$, and $\bar{R}_f$ represent the average rates of return for ith security (portfolio), market rates of return, and risk-free rate, respectively.

Regression results in terms of Eq. 21.12 for both JNJ and MRK are presented in Fig. 21.10. From the estimated α_i, we conclude that JNJ perform worse than MRK during the period of 1990–2009 since the JM of JNJ is smaller than that of MRK.

Interrelationship Among Three Performance Measures. It should be noted that all three performance measures are interrelated. For instance, if $\rho_{im} = \sigma_{im}/\sigma_i\sigma_m = 1$, then the Jensen measure divided by σ_i becomes equivalent to the Sharpe measure. Since

$$\beta_i = \sigma_{im}/\sigma_m^2 \qquad \text{and} \qquad \rho_{im} = \sigma_{im}/\sigma_i\sigma_m$$

the Jensen measure (JM) must be multiplied by $1/\sigma_i$ in order to derive the equivalent Sharpe measure:

[9] The derivation and justification of Jensen investment performance measure can be found in Michael C. Jensen (1968), "The Performance of Mutual Fund in the Period 1945–1964," *Journal of Finance 23*, 389–416.

Data Display

Row	JNJ	MRK	S&P	T-Bill rate
1	0.230108	0.185441	0.036396	0.0749
2	0.616842	0.878595	0.124301	0.0538
3	-0.551293	-0.733803	0.105162	0.0343
4	-0.091578	-0.182990	0.085799	0.0300
5	0.244915	0.142442	0.019960	0.0425
6	0.584558	0.753915	0.176578	0.0549
7	-0.409758	0.235280	0.237724	0.0501
8	0.340804	0.352548	0.302655	0.0506
9	0.287688	0.409696	0.242801	0.0478
10	0.124207	-0.537078	0.222782	0.0464
11	0.139719	0.411867	0.075256	0.0582
12	-0.431191	-0.357447	-0.163282	0.0339
13	-0.078010	-0.013313	-0.167680	0.0160
14	-0.021172	-0.158294	-0.028885	0.0101
15	0.248595	-0.271964	0.171379	0.0137
16	-0.032496	0.036942	0.067731	0.0315
17	0.122480	0.418327	0.085510	0.0473
18	0.034602	0.367425	0.127230	0.0435
19	-0.076435	-0.450781	-0.174081	0.0137
20	0.108473	0.254065	-0.222935	0.0015

Regression Analysis: JNJ-TBill versus S&P-TBill

The regression equation is

JNJ-TBill = 0.0155 + 0.574 S&P-TBill

Predictor	Coef	SE Coef	T	P
Constant	0.01547	0.06708	0.23	0.820
S&P-TBill	0.5739	0.4788	1.20	0.246

S = 0.293711 R-Sq = 7.4% R-Sq(adj) = 2.2%

Fig. 21.10 Regression results of Eq. 21.12 for JNJ and MRK

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.12390	0.12390	1.44	0.246
Residual Error	18	1.55279	0.08627		
Total	19	1.67670			

Unusual Observations

Obs	S&P-TBill	JNJ-TBill	Fit	SE Fit	Residual	St Resid
3	0.071	-0.5856	0.0561	0.0687	-0.6417	-2.25R
7	0.188	-0.4599	0.1231	0.1006	-0.5830	-2.11R

R denotes an observation with a large standardized residual.
Durbin-Watson statistic = 2.52893

Regression Analysis: MRK-TBill versus S&P-TBill

The regression equation is
MRK-TBill = 0.0309 + 0.646 S&P-TBill

Predictor	Coef	SE Coef	T	P
Constant	0.03092	0.09522	0.32	0.749
S&P-TBill	0.6457	0.6797	0.95	0.355

S = 0.416938 R-Sq = 4.8% R-Sq(adj) = 0.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.1569	0.1569	0.90	0.355
Residual Error	18	3.1291	0.1738		
Total	19	3.2859			

Unusual Observations

Obs	S&P-TBill	MRK-TBill	Fit	SE Fit	Residual	St Resid
3	0.071	-0.7681	0.0767	0.0976	-0.8448	-2.08R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 2.47470

Fig. 21.10 (continued)

Table 21.15 Means and standard deviation

State of economy	Probability, P_i	Return R_i (%)	R_iP_i (%)
Project X			
Prosperity	.25	25	6.25
Normal	.50	15	7.50
Recession	.25	5	1.25
	1.00		15.00
Standard deviation $= \sigma_X = 7.07$ %			
Project Y			
Prosperity	.25	40	10
Normal	.50	15	7.5
Recession	.25	−10	−2.5
	1.00		15.00
Standard deviation $= \sigma_Y = 17.68$ %			

$$\begin{aligned}\frac{JM_i}{\sigma_i} &= \frac{\left[\bar{R}_i - R_f\right]}{\sigma_i} - \frac{\left[\overline{R_m} - R_f\right]}{\sigma_m}\frac{(\sigma_{im})}{\sigma_m\sigma_i} \\ &= \frac{\left[\bar{R}_i - R_f\right]}{\sigma_i} - \frac{\left[\overline{R_m} - R_f\right]}{\sigma_m}(\text{common constant}) \\ &= SM_i - SM_m\ (\text{common constant})\end{aligned}$$

If the Jensen measure (JM) is divided by β_i, it is equivalent to the Treynor measure (TM) plus some constant common to all portfolios:

$$\begin{aligned}\frac{JM_i}{\beta_i} &= \frac{\left[\bar{R}_i - R_f\right]}{\beta_i} - \frac{\left[\overline{R_m} - R_f\right]\beta_i}{\beta_i} \\ &= TM_i - \left[\overline{R_m} - \overline{R_f}\right]\end{aligned}$$

21.8 The Mean and Variance Method for Capital Budgeting Decisions

The capital budgeting decision is the manager's decision to undertake a certain project instead of other projects.[10]

Capital budgeting frequently incorporates the concept of probability theory. Consider two projects (project X and project Y) and three states of the economy for any given time (prosperity, normal, and recession). For each of these states, a probability of occurrence can be calculated and an estimate made of its return, see Table 21.15. We can calculate the expected returns $\bar{R}$ for projects X and Y as follows:

[10] This section discussed how can we use statistical distribution method to make capital budgeting under uncertainty. Other methods to perform capital budgeting under uncertainty can be found in Lee, A. C, J. C. Lee, and C. F. Lee. *Financial Analysis, Planning and Forecasting: Theory and Application*, 2nd ed. Singapore: World Scientific Publishing Company, 2009.

$$\bar{R} = \sum_{i=1}^{m} R_i P_i \tag{21.14}$$

where R_i is the return for the ith state of nature and P_i is the probability associated with the ith state of nature. Substituting the information given in Table 21.15 into Eq. 21.13, we obtain

$$\bar{R}_X = 6.25\% + 7.50\% + 1.25\% = 15.00\%$$
$$\bar{R}_Y = 10\% + 7.50\% - 2.50\% = 15.00\%$$

The standard deviation for these returns can be found by using

$$\sigma = \sqrt{\sum_{i=1}^{n} (R_i - \bar{R})^2 p_i} \tag{21.15}$$

Substituting into Eq. 21.14 the information from Table 21.15, $\bar{R}_x = .15$ and $\bar{R}_Y = 0.15$, we obtain

$$\sigma_X = \left[(.25 - .15)^2(.25) + (.15 - .15)^2(.50) + (.05 - .15)^2(.25)\right]^{1/2}$$
$$= 7.07\%$$
$$\sigma_Y = \left[(.40 - .15)^2(.25) + (.15 - .15)^2(.50) + (-.10 - .15)^2(.25)\right]^{1/2}$$
$$= 17.68\%$$

The data given in Table 21.15 can be used to draw histograms of both projects (see Fig. 21.11a). If we assume that rates of return R are continuously and normally distributed, then Fig. 21.10a can be drawn approximately as Fig. 21.11b.

The concept of statistical probability distribution can be combined with capital budgeting to derive the *statistical distribution method* for selecting risky investment projects. The expected return for both projects is 15 %, but because project Y has a normal distribution with a wider range of values, it is the riskier project. Project X has a normal distribution with a larger collection of values closer to the 15 % expected rate of return and is therefore more stable.

21.8.1 Statistical Distribution of Cash Flow

Accounting concepts make it possible to define the net cash flows (C_t) as

$$C_t = (\mathrm{CF}_t - d_t - I_t)(1 - \tau_c) + d_t \tag{21.16}$$

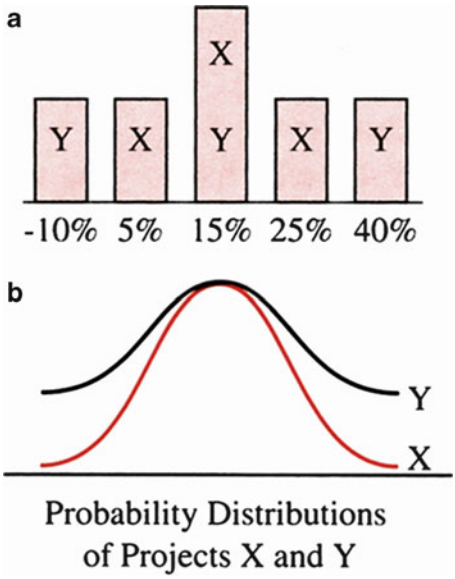


Fig. 21.11 Histograms and probability distributions of projects X and Y

where

$\mathrm{CF}_t = Q_t(P_t - V_t)$
$(CF_t—d_t—I_t)(1—\tau_t) =$ net income
$Q =$ quantity produced and sold
$P =$ price per unit
$V =$ variable costs per unit
$d_t =$ depreciation
$\tau_c =$ tax rate
$I_t =$ interest expense

For this equation, net cash flow is a random number because Q, P, and V are not known with certainty. We can assume that Net C_t has a normal distribution with mean $\bar{C}_t$ and variance σ_t^2, which was similar to that defined in Eq. 7.17.

If two projects have the same expected cash flow, or return, as determined by the expected value defined in Eq. 21.14, we might be indifferent between the projects if we were to make our choice on the basis of return alone. If, however, we also take risk (variance) into account, we will get a more accurate picture of what type of cash flow or return distribution to expect.

With the introduction of risk, a firm is not necessarily indifferent between two investment proposals that are equal in net present value (NPV).[11] We should estimate

[11] See Appendix 3 for the definition of NPV.

both NPV and its standard deviation σ_{NPV} when we perform capital budgeting analysis under uncertainty. Net present value under uncertainty can be defined as

$$\text{NPV} = \sum_{t=1}^{N} \frac{\tilde{C}_t}{(1+R_f)^t} + \frac{S_t}{(1+R_f)^N} - I_0 \tag{21.17}$$

where

$\tilde{C}_t$ = uncertain net cash flow in period t
R_f = risk-free discount rate
S_t = salvage value of facilities
I_o = initial outlay (investment)

The mean of the NPV distribution and its standard deviation can be defined as follows for mutually independent cash flows:

$$\overline{\text{NPV}} = \sum_{t=1}^{N} \frac{\bar{C}_t}{(1+R_f)^t} + \frac{S_t}{(1+R_f)^N} - I_0 \tag{21.18}$$

$$\sigma_{\text{NPV}} = \left(\sum_{t=1}^{n} \frac{\sigma_t^2}{(1+R_f)^{2t}} \right)^{1/2} \tag{21.19}$$

The generalized case for Eq. 21.19 is explored in Appendix 4.

Example 21.13 The Mean and Variance Approach for Capital Budgeting Decisions. A firm is considering the introduction of two new product lines, A and B, that have the same life and have the cash flows, standard deviations of cash flows, and salvage values shown in Table 21.16. Assume a discount rate of 10 %. Both projects have the same expected NPV:

$$\begin{aligned}
\overline{\text{NPV}}_{\text{A}} = \overline{\text{NPV}}_{\text{B}} &= \sum_{t=1}^{5} \frac{\bar{C}_t}{(1+R_f)^t} \\
&= 20\left(\text{PVIF}_{10\%,1}\right) + 20\left(\text{PVIF}_{10\%,2}\right) + 20\left(\text{PVIF}_{10\%,3}\right) \\
&\quad + 20\left(\text{PVIF}_{10\%,4}\right) + 20\left(\text{PVIF}_{10\%,5}\right) - 60 + 5(.6209) \\
&= 20(.9091) + 20(.8264) + 20(.7513) + 20(.6830) \\
&\quad + 20(.6209) - 60 + 5(.6209) \\
&= 18.90
\end{aligned}$$

where PVIF is the present value interest factor (see Appendix 3 for the calculation).

However, because the standard deviation of A's cash flows is greater than that of B's, project A is riskier than project B. This difference can be explicitly evaluated

Table 21.16 Data for Example 21.10 (in thousands)

	Project A	Project B
Initial investment	\$60	\$60
Cash flows		
Year 1	\$20	\$20
Standard deviation	\$4	\$2
Year 2	\$20	\$20
Standard deviation	\$4	\$2
Year 3	\$20	\$20
Standard deviation	\$4	\$2
Year 4	\$20	\$20
Standard deviation	\$4	\$2
Year 5	\$20	\$20
Standard deviation	\$4	\$2
Salvage value	\$5	\$5

only by using the statistical distribution method. To compare the riskiness of the two projects, we calculate the standard deviation of their NPVs. We will assume that cash flows between different periods are perfectly positively correlated. The σ_{NPV} can then be defined (see Appendix 4) as

$$\sigma_{\text{NPV}} = \sum_{t=1}^{N} \frac{\sigma_t}{(1+R_f)^t} \tag{21.20}$$

$$\begin{aligned}
\sigma_{\text{NPV}_\text{A}} &= (\$4)\left(\text{PVIF}_{10\%,1}\right) + (\$4)\left(\text{PVIF}_{10\%,2}\right) + \cdots + (\$4)\left(\text{PVIF}_{10\%,5}\right) \\
&= (4)(.9091) + (4)(.8264) + (.7513)4 + 4(.6830) + 4(.6209) \\
&= 15.16, \text{ or } \$15,160 \\
\sigma_{\text{NPV}_\text{B}} &= (\$2)\left(\text{PVIF}_{10\%,1}\right) + (\$2)\left(\text{PVIF}_{10\%,2}\right) + \cdots + (\$2)\left(\text{PVIF}_{10\%,5}\right) \\
&= (2)(.9091) + (2)(.8264) + (2)(.7513) + (2)(.6230) + (2)(.6209) \\
&= 7.58, \text{ or } \$7,580
\end{aligned}$$

With the same NPV, project B's cash flows would fluctuate \$7,580 per year, and project A's \$15,160. Therefore, B is to be preferred given the same returns, because it is less risky.

21.9 Summary

In this chapter, we examined the concepts and applications of statistical decision theory and saw that it is different from the classical statistics we have worked with in the last 20 chapters. In the context of statistical decision theory, we discussed elements of decision making under uncertainty. Decisions based on extreme values, expected monetary values and utility measurement, Bayes' strategies, and decision trees were explored. In addition, we developed the Von Neumann and Morgenstern

utility and risk-aversion concepts in order to discuss trade-offs between risk and return. The capital asset pricing model and the statistical distribution method for project selection were also investigated.

Questions and Problems

1. What are the basic elements of decision making? Define those elements separately. If we don't know the probabilities for the states of nature, can we still make a decision?
2. John faces the following decision problem:

Study hours per day	High confidence	Average confidence	Low confidence
0	60	40	30
5	80	60	50
10	90	80	60

   With 5 h of study per day, John estimates that there are three different numbers of points he can get on the midterm: 80 with high confidence, 60 with average confidence, and 50 with low confidence. He also has two other possible actions: studying 0 h per day and studying 10 h per day. Estimate the points he can get on the midterm in each of the three states: high, average, and low. Try to use the maximin criterion to choose the best action and specify the most points he can get on the midterm.
3. In question 2, rebuild the table by using the minimax criterion and specify the best action and the best points. Is the best action the same as that of question 2? If yes, is this by chance or is it always true?
4. Reconsider the table in question 2 in the following way:

Study hours per day	High	Average	Low
0	.2	.5	.3
5	.2	.6	.2
10	.4	.5	.1

   If John studies 5 h per day, then he estimates the probabilities in three confidence levels as .2 for high, .6 for average, and .2 for low. The same interpretations apply to the other two levels.

   (a) What are the probabilities for the high level, given 0, 5, and 10 studying hours per day?
   (b) Suppose the probability for each action is $\frac{1}{3}$. What are the probabilities for 0, 5, and 10 h of study per day, given the high level?

5. (a) Using the expected monetary value (EMV) criterion, try to construct a table to determine which action is best.
   (b) 5b. Does applying the EMV criterion yield the same best action as applying the maximin criterion or applying minimax regret criterion?

6. Assume the following utilities for the different midterm points in question 2. The point scores 30, 40, 50, 60, 80, and 90 have utilities of 2, 5, 7.5, 9.5, 11, and 12 units, respectively. Is this a risk-averse, risk-neutral, or risk-loving type of utility function? (The implications of this assessment are worth pondering!)
7. Define the terms *risk-averse, risk-neutral,* and *risk-taking*.
8. (a) Reconstruct an expected utility table by using the table in question 4 and the assumption for utility units in question 6. If John makes his decision in accordance with the criterion of largest expected utility, which action will he choose?
   (b) Redefine *risk-averse, risk*-neutral, and *risk-taking* in terms of the expected utility concept.
9. Given the utility function $U(W) = 10W^{1/2}$, wherein W = payoff,
   (a) Graph the function.
   (b) Does the function exhibit risk aversion? What is your criterion?
   (c) How will changing the constant term 10 to an arbitrary number a affect the answer to parts (a) and (b)? (i.e., by assuming that a is greater or less than 0, what different result will we get?)
10. Mr. Clark has $100 and would like to try his luck in an Atlantic City casino. Suppose he is faced with a 60/40 chance of losing $20 or winning $15. Further suppose that for a fee of $10, he can buy insurance that completely removes the risk.
    (a) If Mr. Clark's utility function is logarithmic for $U(W) = \text{In } W$, is he a risk-averse person? How do you know?
    (b) Will Mr. Clark buy the insurance or take the gamble?
    (c) Say the risk increases to a 70/30 chance of losing $20 or winning $15. How much of a premium will Mr. Clark now pay for insurance to remove the risk completely (assuming he remains risk-averse)?
    (d) Say Mr. Clark's initial wealth increases to $150. What change does this bring about in the risk premium he pays to remove the risk completely (assuming he remains risk-averse)?
11. Lottery A offers a 70 % chance of winning $45 and a 30 % chance of losing $100. Lottery B offers a 60 % chance of winning $55 and a 40 % chance of losing $85. Lottery C offers an 80 % chance of winning $30 and a 20 % chance of losing $110. Without knowing any additional information, such as the utility function, which lottery will you choose? By what criterion?
12. Use the MINITAB and the R_i and R_m information given in the table to calculate systematic risk, which was defined in Eq. 21.5 in the text.

Quarterly rates of return for IBM (R,) and market rates (R_m), second quarter 1981 to second quarter 1991

		Market return	IBM return
1981			
	2	−.03522	−.05835
	3	−.11454	−.04993
	4	.054828	.066697
1982	1	−.08641	.065670
	2	−.02098	.029037
	3	.098622	.224494
	4	.167912	.323475
1983	1	.087599	.066077
	2	.099045	.191154
	3	−.01213	.062993
	4	−.00686	−.03093
1984	1	−.03486	−.05778
	2	−.03769	−.06403
	3	.084345	.185342
	4	.006863	−.00020
1985	1	.080243	.040406
	2	.061939	−.01692
	3	−.05092	.009898
	4	.160369	.264177
1986	1	.130726	−.01864
	2	.049979	−.02574
	3	−.07781	−.07440
	4	.046904	−.09962
1987	1	.204525	.260208
	2	.042166	.089758
	3	.058651	−.06553
	4	−.23232	−.22653
1988	1	.047883	−.05865
	2	.056433	.193728
	3	−.00581	−.08557
	4	.021367	.065872
1989	1	.061752	−.09558
	2	.078373	.036288
	3	.098025	−.01264
	4	.012172	−.12736
1990	1	−.03808	.140345
	2	.053185	.118586
	3	−.14515	−.08438
	4	.078974	.073654
1991	1	.136272	.018451
	2	−.01082	−.13646

13. (a) Given $R_f = 5\ \%$ and $E(R_m) = 10\ \%$, plot the security market line (SML) (write the equation).
    (b) If $\beta_i = 2$ and $E(R_j) = 12\ \%$, will a wise investor *purchase* stock i? Why or why not?
14. Discuss some of the different methods of decision making, and explain when you would use each one.
15. Describe why knowing which outcome you prefer is not adequate for making a choice under uncertainty.
16. Describe why good decisions sometimes result in bad outcomes.
17. What is the maximin criterion? When is it best to use the maximin criterion?
18. What is the minimax regret criterion? When is using this criterion best?
19. Using Example 21.10 in the text as an example, explain how Bayesian analysis can be applied in decision making.
20. What is the expected monetary value (EMV) criterion? Briefly explain how this criterion is applied.
21. Suppose you are interested in evaluating a stock's price. You have analyzed the probability that the stock will go up on any given day as 1/3, the probability that the stock will go down on any given day as 1/3, and the probability that the stock's price will not change on any given day as 1/3. Use a decision tree to show the possible stock price movements for 3 days.
22. Briefly explain what the dominance principle is and how it can be used in risk-and-return analysis.
23. Draw the capital market line. Write down the equation for the capital market line. Explain what the capital market line tells us.
24. What is the capital asset pricing model (CAPM)? What are the assumptions of this model? What does it tell us?
25. What is the CML? What is the SML? How are they similar? How are they different?
26. An investor wants to choose among three investment alternatives: a passbook savings account, a government bond fund, and a growth stock fund. The payoffs for a $20,000 investment are given in the following table.

	State of nature		
Investment	Low growth	Normal growth	High growth
Savings account	$1,000	$1,000	$1,000
Bond fund	$1,500	$1,000	$800
Stock fund	$500	$1,200	$1,500

    (a) Which investment does applying the minimax regret criterion instruct us to choose?
    (b) Which investment does applying the maximin criterion instruct us to choose?

27. Now suppose the investor in question 26 assigns probabilities of .3 to low growth, .4 to normal growth, and .3 to high growth. Use the expected monetary value criterion to determine which investment should be chosen.
28. You are given the following information on the market and on XYZ Company's stock.

    Return on market = 10 %
    Risk-free interest rate = 6 %
    Beta for XYZ stock =1.5

    Compute the expected return on XYZ's stock.
29. You are trying to decide whether you should study a lot or a little for your statistics midterm. You construct the following grade-payoff table.

	State of nature	
Action	Easy test	Hard test
Study a lot	98	95
Don't study	90	55

    (a) Use the minimax regret criterion to determine how much to study.
    (b) Use the maximin criterion to determine how much to study.

30. A studio that has just produced a new movie must decide when to release it. The possible actions are

    A_1: release the movie in the spring
    A_2: release the movie in the fall
    A_3: release the movie at Christmas time
    A_4: release the movie in the summer

    The possible states of nature are

    S_1: low movie attendance
    S_2: average movie attendance
    S_3: high movie attendance

    The payoff table is

	State of nature		
Action	S_1	S_2	S_3
A_1	−20	10	20
A_2	10	10	10
A_3	25	35	45
A_4	20	19	40

(a) Which of these actions will the studio choose if it uses the minimax criterion?
(b) Which of these actions will the studio choose if it uses the maximin criterion?

31. Suppose that in question 30, you have assigned the probabilities of .2 to low movie attendance, .5 to average movie attendance, and .3 to high movie attendance. Use the expected monetary value criterion to determine when the studio should release the movie.
32. What is a decision tree? Briefly explain how a decision tree can be used in decision theory.
33. Consider the following payoff table, where the cell entries are in dollars.

	Alternative			
Outcome	A	B	C	D
1	5	7	5	4
2	3	2	2	7
3	9	8	7	5
4	7	10	6	4

Can any alternatives be eliminated by using dominance?

34. Use an example to show the similarities and differences between the Sharpe and the Treynor investment performance measures.
35. A local deli prepares fresh potato salad for its customers every day. The unsold salad has to be thrown away. The demand for potato salad can be classified as low (100 lb), medium (200 lb), or high (300 lb). The production runs being considered are 100, 200, and 300 lb. The payoffs for all combinations of production and demand are shown here.

	Demand		
Production	100	200	300
100	400	300	−100
200	−400	600	700
300	−800	−300	1,500

(a) What is the maximin solution of this problem?
(b) What is the expected monetary value of each action if the probabilities of demand being low, medium, and high are .3, .3, and .4, respectively?

Use the following information to answer questions 36–40. A manufacturer is planning its production for the next 6 months. It has to decide how much of an important ingredient to keep in inventory. The demand for the ingredient may be low, medium, or high. The manufacturer is considering holding either a low or a high amount of inventory. The possible payoffs for all the combinations of inventory holding and demand are shown here.

	Demand		
Inventory holding	Low S_1	Medium S_2	High S_3
Low A_1	200	300	300
High A_2	100	200	500
Probability	.3	.2	.5

36. What is the minimax regret solution to this problem?
37. What is the expected payoff of each action?
38. An economics consultant predicts that the demand for the ingredient will be low in the next 6 months. In the past few years, this economist has provided forecasting about the demand for the ingredient. The track record of the consultant is summarized by the following conditional probability distribution:

$$\begin{array}{lll} p(H|S_1) = .6 & p(H|S_2) = .4 & p(H|S_3) = .1 \\ p(M|S_1) = .2 & p(M|S_2) = .2 & p(M|S_3) = .1 \\ p(L|S_1) = .2 & p(L|S_2) = .5 & p(L|S_3) = .8 \end{array}$$

Assume that this time, the economist predicts a low demand for the future. Find the posterior distribution of S_1, of S_2, and of S_3.

39. Use the foregoing information to evaluate the action of accumulating a high inventory. (Obtain the expected monetary value by using posterior probability.)
40. Write out the decision tree for this question.

Use the following information to answer questions 41–45. A company is considering what size copying machine it should lease. The copier comes in three different sizes: small, medium, and large. A larger machine can handle more work, but it also costs more. The demand for the machine in the next year is uncertain. The cost of leasing a smaller machine is lower than the cost of leasing a larger one. However, at those times when the small machine could not handle the high demand, the company would have to lease a second machine and pay a significantly higher cost than if it had leased a larger machine in the first place. The possible costs of leasing the three different copiers under conditions of low and high demand are presented in the following table.

	Future demand	
Size of copier	Low (S_1)	High(S_2)
Small	400	800
Medium	500	900
Large	600	600
Pr(S)	.4	.6

41. Find the maximin solution.
42. Find the minimax regret solution.
43. Would you lease a medium-sized machine under any circumstances? Why or why not?
44. What are the expected costs of leasing a large machine?

45. An economic consultant predicts that the demand for the machine will be high the next year and suggests that the company should therefore lease a large machine. Assume that the consultant has the following track record of predicting demand.

$$\Pr(I_1|S_1) = .8 \qquad \Pr(I_2|S_1) = .2$$
$$\Pr(I_1|S_2) = .2 \qquad \Pr(I_2|S_2) = .8$$

where I_1 indicates that the consultant predicts S_1 and I_2 indicates that the consultant predicts S_2. Do you agree with the consultant's advice?

46. A limousine chauffeur is going to take a guest from the hotel to the airport. To catch the flight, the chauffeur has to arrive at the airport in 30 min. There are two routes to the airport: a local route and the highway. The chauffeur has found that it always takes him 25 min to get to the airport when he takes the local route. When he takes the highway, the time consumed depends on the traffic. When the highway is jammed, it takes him 36 min to get to the airport. When the highway is clear, the trip takes only 10 min. There is a.10 probability that the highway will be jammed.

    (a) Which route should the chauffeur take on the way to the airport?
    (b) Which way should the chauffeur take when he is coming back to the hotel?

47. Does the risk-averse decision maker ever take any risk?

Use the following information to answer questions 48–50. Assume an investor has the utility function $W^{1/3}$, where W is wealth. The state government issues a lottery ticket that pays the winner \$300. The lottery ticket costs \$1. The chance of winning the lottery is 1/200. The investor has \$270 in original wealth.

48. What is the expected value of this lottery? What is the investor's expected utility if he buys the lottery?
49. Use the lottery case to show that the investor is risk-averse.
50. Will the investor buy the lottery ticket if his utility function is W^3?
51. The owner of a personal computer company is considering whether to install a large or a small new assembly line. The possible payoffs (in thousands of dollars) depend on the state of the economy and are presented in the following table.

	State	
Size of assembly line	Boom	Recession
Large	20	5
Small	10	8
Probability	.3	.7

    (a) Find the minimax regret solution.
    (b) Find the expected payoff of installing a large assembly line when the probability of boom and the probability of recession are .3 and .7, respectively.
    (c) If the beginning wealth of the company is 20 and the utility function of the owner is $W^{1/2}$, what is the expected utility of the two options?

52. Mr. Montero is deciding how to invest the money for his son's tuition, which is due 2 months from today. He can put the money in a 2-month certificate of deposit (CD) earning a 10 % annual rate of interest, or he can put the money in a 1-month CD now and earn 9 %. If he puts the money in the 1-month CD, then a month later he will have to invest it in another 1-month CD. The rate 1 month from today is uncertain. Both CDs are protected by FDIC insurance. Assume that Mr. Montero knows the 1-month rate in the next month follows a normal distribution with a mean of 11 % and a standard deviation of 2 %. What will be his choice if he is a risk averter? What will be his choice if he is risk-neutral?

Use the following information to answer questions 53–55. Ms. Jones is thinking of investing in a new project that will cost $1,000 to start. There are two ways to raise this $1,000. She can take out $1,000 from her own pocket or take out $500 and invite a friend to share the investment. The investment will generate the following revenues, depending on the outcome of the investment.

	State 1	State 2	State 3
Revenue	$800	$1,000	$1,500

The revenue will be equally split between Ms. Jones and her friend if they share the project. It is estimated that the probabilities of states 1, 2, and 3 are 1/3, 1/3, and 1/3, respectively.

53. (a) What is the expected net gain of the project if Ms. Jones undertakes the investment alone?
    (b) What is the expected net gain of the project for Ms. Jones if the investment is shared?

54. Ms. Jones hired Dr. Lee, an economics consultant, to evaluate the probabilities of states 1, 2, and 3. Suppose this consultant has the following track record.

	Actuality, $P(I_i \mid S_i)$		
	S_1	S_2	S_3
I_1	.8	.2	.2
I_2	.1	.6	.2
I_3	.1	.2	.6

    (a) Obtain the posterior distribution when Dr. Lee predicts I_1.
    (b) Evaluate the expected payoff for Ms. Jones of sharing the project.

55. Ms. Jones has the utility function $U = f(W) = W^{1/2}$, and her initial wealth is $1,000. Should she invest in the project? If so, should she invest alone? Use the original probability function to answer this question.

56. The owner of the New Land Food Corporation is considering a new project that has the following possible payoffs (in thousands of dollars).

Profits	Probability	Profits	Probability
200	.1	0	.15
1,000	.25	−5,000	.2
5,000	.2	−1,000	.1

The owner's current assets are worth about \$5,000. His utility function is $U = W^{1/2}$.

(a) What is the expected value of this project?
(b) What is the expected utility of this project?

57. The owner of North American toy company is considering enlarging its production capacity to meet increasing future demand. The company can either expand its old plant or establish a new plant. The possible payoffs of these two actions are related to the increase in future demand and are shown in the following table.

		Demand (in thousands of dollars)	
Action	Low	Medium	High
Expansion	150	250	250
New plant	0	250	500
Probability	1/3	1/3	1/3

(a) Write out the decision tree, and determine which action the owner should take if he uses the minimax regret approach.
(b) Assume the net worth of the firm at this stage is \$500 thousands. The utility function of the owner is $U(W) = 400 - 4000/W^{1/2}$. Which action will generate the higher expected utility?

58. Ace Corporation is sending a shipment of crystal balls from North Carolina to California. There is a chance that the shipment may be damaged, so Ace Corporation is thinking of buying insurance to cover the shipment. The possible costs that buying and not buying insurance may entail are presented in the following table.

	Shipment wrecked	Shipment safe
Buying	\$100	\$100
Not buying	\$5,000	0
Probability	.01	.99

What is the expected value of the insurance policy? If the insurance policy is not worth the cost of insurance (\$100), why do people buy it?

59. An ice cream stand at the beach wants to order some ice cream for the coming weekend. The demand for ice cream depends on the weather. The ice cream has to be ordered in 50-gallon units.
The profits that selling ice cream yield under different combinations of state, and ice cream order are presented here.

	Bad weather S_1	Good weather S_2
50 gallons	\$500	\$500
100 gallons	\$300	\$900

Historically, the probabilities of S_1 and S_2 during this time of the year are about .4 and .6, respectively.

(a) Find the minimax regret solution.
(b) Find the highest-expected-value solution.
(c) Compare the utility of ordering 50 gallons versus 100 gallons if the ice cream stand owner's utility function is

$$U = f(E, S) = 50E - 25S$$

where E is the expected wealth and S is the standard deviation of wealth.

Use the following information to answer questions 60–63. A new company is formed to invest in a new project. This company is going to raise the needed capital, $100,000, by issuing $50,000 bonds and $50,000 stock. The bondholder is guaranteed a 10 % interest rate regardless of the performance of the company. The stockholder will receive whatever is left after bondholders are paid. An investor is thinking of investing $40,000 in the company for 1 year. A year later, she will pull out of the investment. She can put the money in any combination of bonds and stock. The possible payoffs of the project (in thousands of dollars) are recorded here.

	Recession S_1	Stable economy S_2	Boom S_3
Earnings before interest	5	10	20
Interest	5	5	5
Earnings after interest	0	5	15
Value of all stocks	50	55	65
Probability	.4	.3	.3

60. The investor can choose to put all of her $40,000 in either bonds or stock. What is the expected value for each of these two options at the end of the year? What is the standard deviation of these two options?
61. Assume that the investor puts her $40,000 in a portfolio consisting x percent of bonds and $(1 - x)$ percent of stock. What is the expected value of this portfolio?
62. Assume that the investor's initial wealth is $40,000 and that her utility function is $U = W^{1/2}$. What is the expected utility of the portfolio described in question 61?
63. A stock analyst has just released a report saying that the economy will be good in the coming year. His track record is

$\Pr(\text{good}\|S_1) = .2$	$\Pr(\text{bad}\|S_1) = .8$
$\Pr(\text{good}\|S_2) = .5$	$\Pr(\text{bad}\|S_2) = .5$
$\Pr(\text{good}\|S_3) = .8$	$\Pr(\text{bad}\|S_3) = .2$

Evaluate the portfolio made up entirely of stock and that made up entirely of bonds.

64. President Reagan was interested in establishing a "defensive wall" to blunt the threat of a nuclear missile attack against this country. The plan was commonly known as "Star Wars." Assume a defense contractor comes up with two defense systems. System A will destroy 60 % of the missiles launched against this country, but 40 % of the missiles will get through. System B has a 60 % probability of destroying all the missiles launched but a 40 % chance of letting all the missiles get through. Should you use the expected value approach to choose which system to support? Make your choice using common sense.
65. Peter Campbell plans to invest in real estate income property. He plans to hold that property for 7 years. He is considering two income properties, A and B. Their initial investment, cash flows, standard deviations of cash flows, and resale values are as follows.

Property A (in thousands of dollars)

Year	Cash Flow	Standard deviation
0	−60	0
1	18	1
2	18	1.1
3	18	1.3
4	18	1.4
5	18	1.5
6	18	1.7
7	18	1.9
7	400 (resale value)	0

Property B (in thousands of dollars)

Year	Cash Flow	Standard deviation
0	−60	0
1	18	.9
2	18	1.1
3	18	1.4
4	18	1.5
5	18	1.6
6	18	1.8
7	18	1.9
7	400 (resale value)	0

Use Microsoft Excel to calculate the NPV and σ_{NPV} of each property. (Assume a discount rate of 12 %.) According to the results, which property should Peter choose?

66. The MINITAB output on pages 1113–1114 is a market model for Ford stock return. Please explain the result in accordance with Eqs. 21.5 and 21.6. (Hint: Refer to Example 21.11.)
67. Assuming that 2 conditions in Table 21.17 change as follows: (1) test cost changes to $25,000, and (2) under the condition that the test is positive, the probability of successful drilling is .7 (failure is .3), use the methods introduced in Appendix 1 to analyze the oil drilling problem again.

68. Sandy is going to make investment for the $10,000 which she deposited for the past 2 year. Her financial advisor presented the following tables which show (1) expected profits (in $10,000's) for various states of nature and their probabilities and (2) the advisor's prediction ability about the state of nature.

	State of nature			
	Recession S_1	Flat S_2	Boom S_3	
Probability of nature	0.2	0.5	0.3	EMV
Bonds	−5.0	1.0	10.0	2.5
Stocks	−30.0	3.0	20.0	1.5

		State of nature		
		Recession S_1	Flat S_2	Boom S_3
Advisor's	Recession Z_1	0.7	0.2	0.1
Prediction	Flat Z_2	0.2	0.6	0.1
	Boom Z_3	0.1	0.2	0.8

Revise the prior probabilities in light of advisor's prediction ability.

69. (Problem 68 continued.) If the financial advisor predicts a boom state, what is the EMV of the investing bond, using the revised probability?
70. (Problem 68 continued.) If the advisor predicts a recession state, what is the EMV of the investing bond, using the revised probability?
71. (Problem 68 continued.) If the advisor predicts a flat state, what is the EMV of investing bond, using the revised probability?
72. (Problem 68 continued.) Compute the EMV of investing stock, using the revised probability under each predicted state.
73. (Problem 68 continued.) What is the optimal investment plan based on the advisor's prediction state?

MINITAB Output of Market Model for Ford (for Question 66)

```
MTB > PRINT C2-C4
```

Data Display

ROW	Year	Ford	Market
1	70	0.4260	0.0010
2	71	0.2933	0.1080
3	72	0.1717	0.1557
4	73	−0.4512	−0.1737
5	74	−0.0968	−0.2964
6	75	0.3960	0.3149
7	76	0.4614	0.1913
8	77	−0.2067	−0.1153
9	78	−0.0026	0.0105
10	75	−0.1479	0.1223
11	80	−0.2938	0.2586
12	81	−0.1025	−0.0994

(continued)

ROW	Year	Ford	Market
13	82	1.3212	0.1549
14	S3	0.1286	0.1706
15	34	0.0980	0.0115
16	35	0.3237	0.2633
17	86	0.0081	0.1462
18	37	0.3961	0.0203
19	38	−0.2995	0.1240
20	89	−0.0767	0.2725
21	90	−0.3209	−0.0656

```
MTB > BRIEF 2
MTB > REGRESS C3 1 C4;
SUBC> DW.
```

Regression Analysis

```
The regression equation is
Ford = 0.0306 + 0.878 Market
```

Predictor	Coef	StDev	T	P
Constant	0.03057	0.09090	0.34	0.740
Market	0.8777	0.5234	1.66	0.110

```
S = 0.3756 R-Sq = 12.9% R-Sq(adj) = 8.3%
Analysis of Variance
Source DF SS MS F P
Regression 1 0.3968 0.3968 2.81 0.110
Residual Error 19 2.6309 0.1411
Total 20 3.0777
Unusual Observations
Obs Market Ford Fit StDev Fit Residual St Resid
 5 -0.296 -0.0968 -0.2296 0.2110 0.1328 0.43 X
 13 0.155 1.3212 0.1665 0.0920 1.1547 3.17R
R denotes an observation with a large standardized
residual
X denotes an observation whose X value gives it large
influence.
Durbin-Watson statistic = 1.30
```

Project V: Project for Selected Topics in Statistical Analysis

1. Test the randomness of the annual rates of return for the Dow Jones' 30 firms which have been calculated in Project IV.
2. Use Microsoft Excel and the Holt-Winters Model to forecast the annual EPS and DPS for JNJ, IBM and AT&T in accordance with the data in Project IV.
3. Use the data obtained in Project IV and the annual 3-month T-Bill rate presented in Table 2.1 to estimate the Jensen investment performance for all the Dow Jones' 30 firms.
4. Use the data in Project IV to estimate both Sharpe and Treynor Investment Performance measures for Dow Jones' 30 firms.
5. Draw implications from the estimate of Jensen, Treynor and Sharpe performance measures.
6. Use the statistical results obtained in Projects I-V to write a summary executive report for your boss.

Download monthly adjusted close price data of JNJ and S&P 500 index from Yahoo Finance during the period from January 2005 to current month to do the following:

7. Calculate the monthly rates of return for JNJ and S&P 500 index and test the randomness of the monthly rates of return for JNJ.
8. Go to St. Louis Federal Reserve Bank website as the link below to download the monthly data of 3-month T-Bill rate: http://research.stlouisfed.org/fred2/data/TB3MS.txt and estimate Jensen investment performance for JNJ. (Remark: the format of the downloaded data is annual rate, therefore the monthly 3-month T-bill rate is the original data divided by 12)
9. Use the monthly rates of return and 3-month T-bill rate data to estimate both Sharpe and Treynor Investment Performance measures for JNJ.

Appendix 1: Using the Spreadsheet in Decision-Tree Analysis

J. M. Jones (1986, *European Journal of Operation Research,* pp. 385–400) showed how the Lotus 1-2-3 spreadsheet package can be used to construct an entire decision tree. Using the information presented in Fig. 21.5, we use Lotus 1-2-3 to construct the decision tree in Fig. 21.12. In this figure, D represents the decision node and C represents the event (chance) node, which correspond to □ and ○ in Fig. 21.5, respectively. Figure 21.12 illustrates all the information we have discussed so far in a more systematic fashion.

There are three steps in applying a spreadsheet to decision-tree analysis. Data from Example 21.8 is used to show how these three steps can be executed.

1. Building the Decision Tree on the Spreadsheet (Lotus 1-2-3)
   In the Lotus 1-2-3 spreadsheet, we know that the cells contain either numbers or labels; we can build the tree on spreadsheet by adopting the following conventions:

   (a) Denote decision nodes by D.
   (b) Denote chance nodes by ©.
   (c) Denote the decision emanating from decision nodes and chance outcomes emanating from chance nodes by appropriate labels.
   (d) Provide a connective structure for the tree using vertical and horizontal dashed line segments.

2. Solving the Tree
   The two main tasks involved in solving the decision tree are averaging out and folding back. Because the Lotus 1-2-3 spreadsheet has excellent computational abilities, these two tasks can be done easily. The process is as follows:

   (a) Create a "master table" within the spreadsheet that incorporates all of the values used as input in the process of developing the tree (see Table 21.17)
   (b) Calculate all yields in the tips of the tree from the master table values. (These yields are the net of all costs involved.)
   (c) Calculate all probabilities needed in the tree from the corresponding values in the master table and put them in the appropriate places of the tree.
   (d) Use the built-in calculating capabilities of the spreadsheet program to perform the averaging out and folding back process.

3. Sensitive Analysis
   The input values in the master table may be subject to change. We can change the values in the master table and then get results under different situations. For example, if we change the drilling cost from $100,000 to $50,000, the overall yield changes from $200,000 to $225,000. The new decision tree is shown in Fig. 21.13.

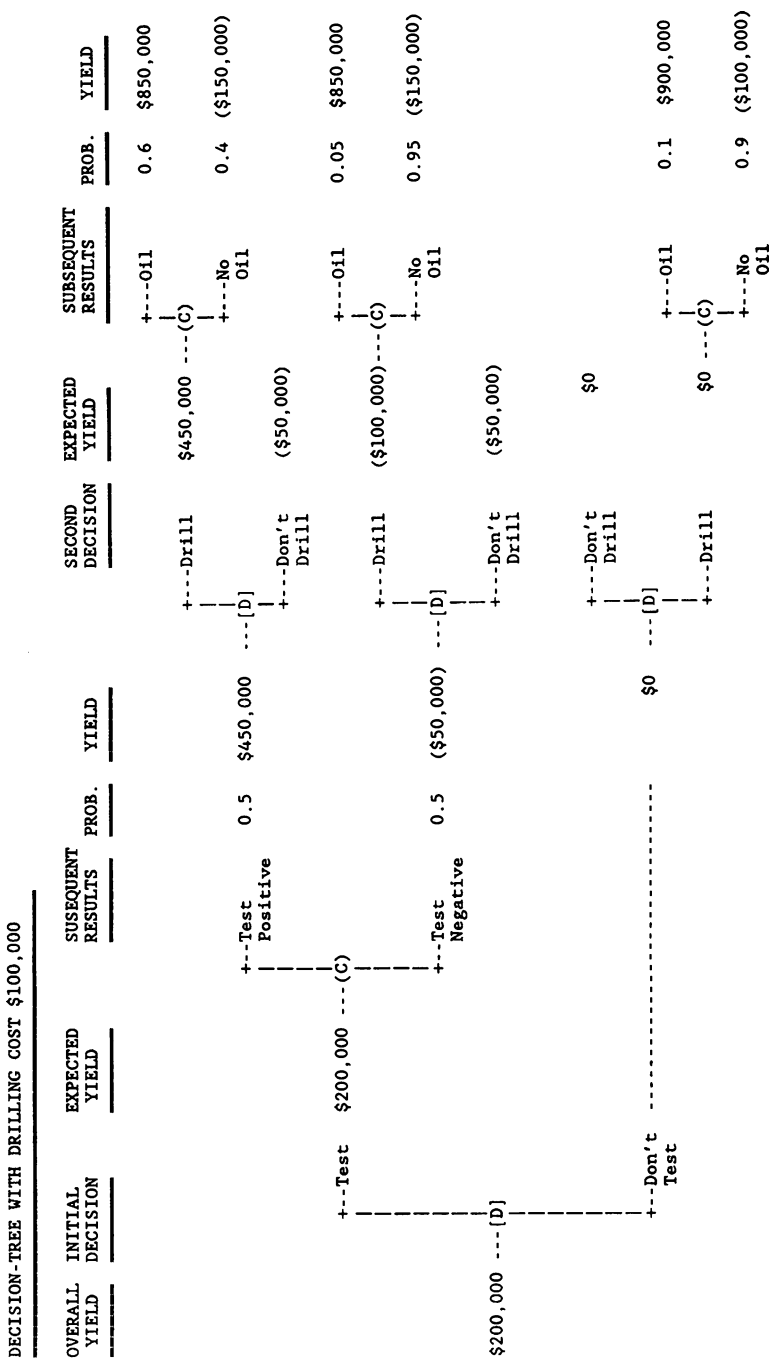


Fig. 21.12 Decision tree with drilling cost of $ 100,000

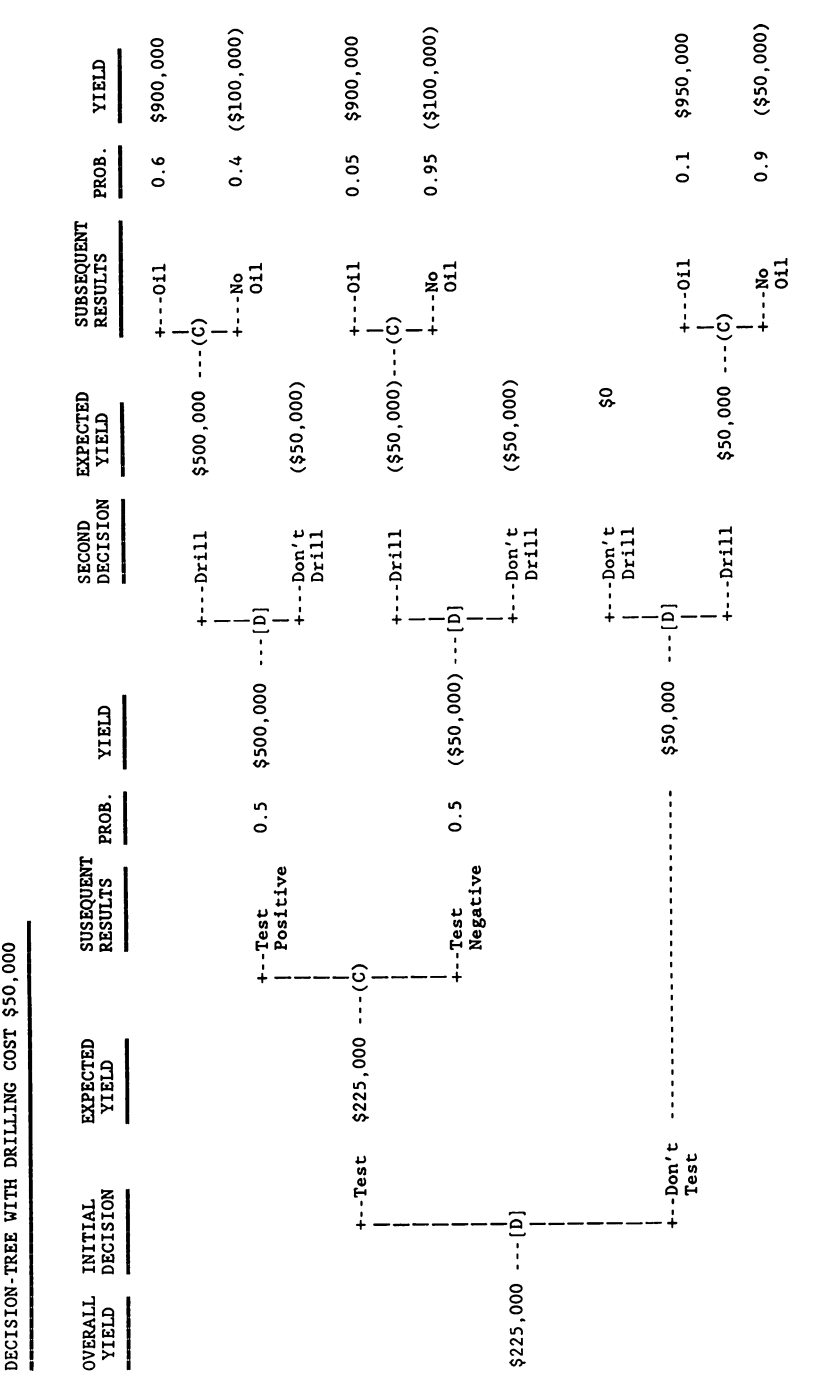


Fig. 21.13 Decision tree with drilling cost of $50,000

Table 21.17 Numbers for the oil drilling problems

Test cost		$50,000	
Drilling cost		$100,000	
Payoff for successful drilling		$1,000,000	
Probability for test results			
Positive	.5		
Negative	.5		
Probability for drilling results			
Test positive		No test	
Success	.6	Success	.1
Fail	.4	Fail	.9
Test negative			
Success	.05		
Fail	.95		

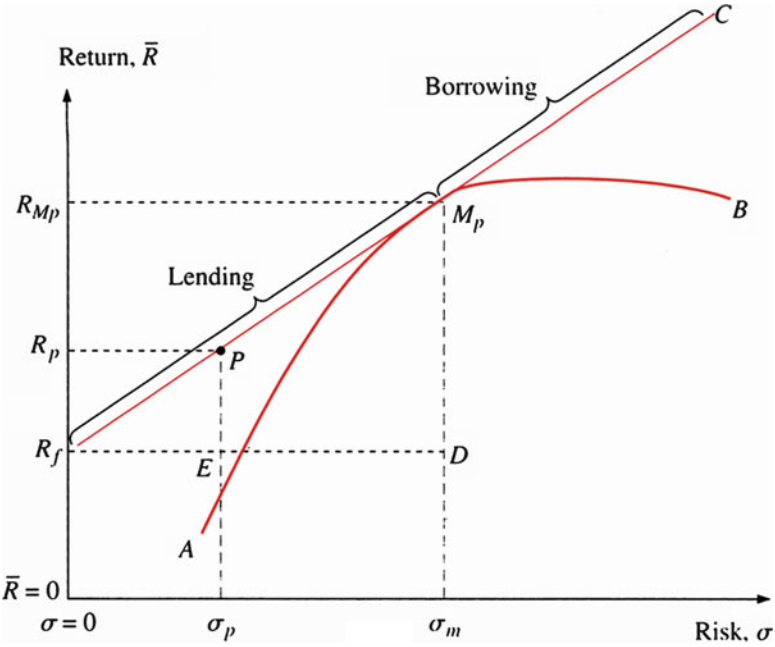


Fig. 21.14 The capital market line

Appendix 2: Graphical Derivation of the Capital Market Line

The term *risk-free assets* in general refers to government securities such as Treasury bills (T-bills). These assets are backed by the federal government and are default-free. In other words, T-bills are riskless; the cash flow from them is certain. An investor's portfolio can be composed of different sets of portfolio opportunities, which may include risk-free assets with a return of R_f, shown on the vertical axis of the risk and return space in Fig. 21.14.

The opportunity to invest in risk-free assets that yield a return of R_f frees the investor to create portfolio combinations that include some risky assets. The investor is able to achieve any combination of risk and return that lies along the line connecting R_f and a point tangent to M_p, the *market portfolio*. All the portfolios along the line R_fM_PC are preferred to the risky portfolio opportunities on the curve AM_pB, because they all have higher expected returns and some risk. Therefore, the points of the line R_fM_pC represent the best attainable combinations of risk and return.

At point R_f, an investor has all available funds invested in the riskless asset and expects to receive the return *of* R_f. The portfolios along the line R_fM_P contain combinations of investments in the risk-free asset and investments in a portfolio of risky assets, M_p. In a sense, the investors who hold these portfolios lend the government money at the risk-free rate R_f —hence the name *lending portfolio*.

At point M_p, the investor holds only risky assets, having put all her wealth into the risky asset, or market, portfolio. At M_p, investors receive a rate of return R_m and undertake risk σ_m.

If it is assumed that investors can borrow money at the risk-free rate R_f and invest this money in the risky portfolio M_p, they will be able to construct portfolios with higher rates of return but higher risks along the line extending beyond M_p. The extent of movement along the line M_PC is regulated by the margin requirements imposed by the various branches of government. The margin requirements stipulate the minimum amount of money investors must pay to buy stock. The higher the margin requirement, the shorter the line M_PC. The amount of money that investors can borrow may also be limited by the creditworthiness of the borrowers. The portfolios along segment M_PC are called *borrowing portfolios*, because the investor must borrow funds in order to achieve these combinations of risk and return. The new efficient frontier becomes R_fM_pC and is referred to as the capital market line (CML). The capital market line describes the relationship between expected return and total risk.

Equation 21.7 can be derived geometrically. An investor has three choices in terms of investments. She may invest in R_f, the riskless asset; in the market portfolio M_p; or in any other efficient portfolio along the efficient frontier, such as portfolio P in Fig. 21.14.

If the investor puts her money into the riskless asset, she can receive a return of R_f. If she invests in the market portfolio, she can expect an average return of R_m and risk of σ_m. If she invests in portfolio P, she can expect an average return of R_p with risk of σ_p. The difference between R_m and R_f $(R_m - R_f)$ is called the *market risk premium.*

The investor in portfolio P takes on a risk of σ_P; her risk premium is $(R_p - R_f)$, which is less than the risk of an investor who holds portfolio M_p.

By geometric theory, triangles R_fPE and R_fM_pD are similar—that is, they are directly proportional. Therefore,[12]

$$E(R_p) - R_f = [E(R_m) - R_f] \frac{\sigma_p}{\sigma_m} \tag{21.21}$$

Appendix 3: Present Value and Net Present Value

In this appendix, we review the concepts of present value and net present value.

Present Value

Because many investment projects will generate returns for several years into the future, it is important to assess the present (current) value of future payments. Suppose a payment is to be received in t years' time and the risk-free annual interest rate for a period of t years is r_t. We know that the future value at the end of t years is $(1 + r_t)^t$ per dollar. Conversely, it follows that the **present value** of a dollar received at the end of t years is

$$\text{Present value per dollar PVIF}(r_t, t) = \frac{1}{(1 + r_t)^t} \tag{21.22}$$

For example, say \$1,000 is to be received in 4 years' time. At an annual interest rate of 8 %, the present value of this future receipt is

$$\frac{1,000}{(1.08)^4} = \$735.03$$

[12] Because $\Delta PR_fE \approx \Delta M_pR_fD$ it follows that

$$\frac{E(R_p) - R_f}{E(R_m) - R_f} = \frac{\sigma_p}{\sigma_m}$$

from this equation, we obtain Eq. 21.21.

Net Present Value

More generally, we can consider a stream of annual receipts, which may be positive or negative. Suppose that, in dollars, we are to receive C_0 now, C_1 in 1 year's time, C_2 in 2 years' time, and so on, until finally we receive C_N in year N. Again, let r_t denote the annual rate of interest for a period of t years. Then, to find the net present value of this stream of receipts, we simply add the individual present values, obtaining

$$\text{NPV} = C_0 + \frac{C_1}{(1+r_1)^1} + \frac{C_2}{(1+r_2)^2} + \cdots + \frac{C_N}{(1+r_N)^N} = \sum_{t=0}^{N} \frac{C_t}{(1+r_t)^t} \quad (21.23)$$

Typically, the rate of interest r_t depends on the period t. When a constant rate, r, is assumed for each period, the *net present value* formula, Eq. 21.23, simplifies to

$$\text{NPV} = \sum_{t=0}^{N} \frac{C_t}{(1+r)^t} \quad (21.24)$$

Example 21.14 NPV Criteria for Capital Budgeting Decisions. A corporation must choose between two projects. Each project requires an immediate investment, and additional costs will be incurred in the next year. The returns from these projects will be spread over a period of 4 years. The following table shows the dollar amounts involved:

		Year 0	Year 1	Year 2	Year 3	Year 4
Project A	Costs	80,000	20,000	0	0	0
	Returns	0	20,000	30,000	50,000	50,000
Project B	Costs	50,000	50,000	0	0	0
	Returns	0	40,000	60,000	30,000	10,000

At first glance, these data might suggest that for project A total returns exceed total costs by \$50,000, whereas the amount of this excess for project B is only \$40,000, signaling a preference for project A. However, such an argument neglects the timing of the returns. See what happens when we calculate the present values of the net receipts for each project, assuming an annual interest rate of 8 % over the period:

		Year 0	Year 1	Year 2	Year 3	Year 4
Project A	Net returns	−80,000	0	30,000	50,000	50,000
	Present values	−80,000	0	25,720	39,692	36,751
Project B	Net returns	−50,000	−10,000	60,000	30,000	10,000
	Present values	−50,000	−9,259	51,440	23,815	7,350

We must compare the sums of these present values when evaluating the projects. For project A, substituting $r = .08$ into Eq. 21.24 yields

$$\begin{aligned}\text{NPV} &= -80{,}000 + \frac{0}{(1.08)^1} + \frac{30{,}000}{(1.08)^2} + \frac{50{,}000}{(1.08)^3} + \frac{50{,}000}{(1.08)^4} \\ &= -80{,}000 + 0 + 25{,}720 + 39{,}692 + 36{,}751 \\ &= \$22{,}163\end{aligned}$$

Similarly, for project B,

$$\begin{aligned}\text{NPV} &= -50{,}000 - \frac{10{,}000}{(1.08)^1} + \frac{60{,}000}{(1.08)^2} + \frac{30{,}000}{(1.08)^3} + \frac{10{,}000}{(1.08)^4} \\ &= -50{,}000 - 9{,}259 + 51{,}440 + 23{,}815 + 7{,}350 \\ &= \$23{,}346\end{aligned}$$

It emerges, then, that if future returns are discounted at an annual rate of 8 %, the net present value is higher for project B than for project A. Project B is preferable because it provides the firm with larger cash flows in the early years, giving the firm a greater opportunity to reinvest the funds.

Appendix 4: Derivation of Standard Deviation for NPV

In Sect. 21.8 we discussed calculation of the standard deviation of NPV where cash flows are perfectly positively correlated or where they are independent of each other. Now we develop a general formula for the standard deviation of NPV for use in all cash flow relationships.

The general equation for the standard deviation of NPV, σ_{NPV}, with a mean of

$$\overline{\text{NPV}} = \sum_{t=1}^{N} \frac{\bar{C}_t}{(1+R_f)^t} + \frac{\bar{S}_t}{(1+R_f)^N} - I_0$$

is

$$\sigma_{\text{NPV}} = \left(\sum_{t=1}^{N} \frac{\sigma_t^2}{(1+R_f)^{2t}} + \sum_{t=1}^{N} \sum_{\tau=1}^{N} W_t W_\tau \text{Cov}(C_\tau, C_t) \right)^{1/2}, \quad \tau \neq t \qquad (21.25)$$

where

σ_t^2 = variance of cash flows in the tth period
W_t, W_τ = discount factor for the tth and the τth period, respectively; that is,

$$W_t = \frac{1}{(1+R_f)^t} \quad \text{and} \quad W_\tau = \frac{1}{(1+R_f)^\tau}$$

$\text{Cov}(C_\tau, C_t)$ = covariability between cash flows C_τ and C_t

Table 21.18 Variance–Covariance Matrix

$W_1^2\sigma_1^2$	$W_1W_2\text{Cov}(C_1,C_2)$	$W_1W_3\text{Cov}(C_1,C_3)$
$W_2W_1\text{Cov}(C_2,C_1)$	$W_2^2\sigma_2^2$	$W_1W_3\text{Cov}(C_2,C_3)$
$W_3W_1\text{Cov}(C_3,C_1)$	$W_2W_3\text{Cov}(C_2,C_3)$	$W_3^2\sigma_3^2$

Cash flows between periods t and τ are generally related. Therefore, $\text{Cov}(C_\tau, Ct)$ is an important factor in the estimation of σ_{NPV}. The magnitude, sign, and degree of the relationships of these cash flows depend on the economic operating conditions and on the nature of the product or service being produced. If there are only three periods, then all terms within the parentheses in Eq. 21.25 can be presented as in Table 21.18. The summation of the diagonal elements $\left[W_1^2\sigma_1^2,\ W_2^2\sigma_2^2,\ W_3^2\sigma_3^2\right]$ of Table 21.18 results in the first part of Eq. 21.25, or

$$\sum_{t=1}^{N} \frac{\sigma_t^2}{\left(1+R_f\right)^{2t}}$$

The summation of all other elements in Table 21.18 gives the second portion of Eq. 21.25, or

$$\sum_{t=1}^{N}\sum_{\tau=1}^{N} W_tW_\tau\text{Cov}(C_\tau, C_t), \quad t \neq \tau$$

Equation 21.25 is the general equation for σ_{NPV}. Both Eq. 21.20 for σ_{NPV} under perfectly positively correlated cash flow and Eq. 21.19 for independent cash flows are special cases derived from the general Eq. 21.25. If $\rho_{12} = \rho_{13} = \rho_{23} = 1$, then $\text{Cov}(C_1,C_2) = \sigma_1\sigma_2$, $\text{Cov}(C_1,C_2) = \sigma_1\sigma_3$, and $\text{Cov}(C_2,C_3) = \sigma_1\sigma_3$. Therefore, Eq. 21.25 reduces to

$$\begin{aligned}\sigma_{\text{NPV}} &= \left(\frac{\sigma_1^2}{\left(1+R_f\right)^2} + \frac{\sigma_2^2}{\left(1+R_f\right)^4} + \frac{\sigma_3^2}{\left(1+R_f\right)^6} + \frac{2\sigma_1\sigma_2}{\left(1+R_f\right)^3} + \frac{2\sigma_1\sigma_3}{\left(1+R_f\right)^4}\right.\\ &\quad \left. + \frac{2\sigma_2\sigma_3}{\left(1+R_f\right)^5}\right)^{1/2}\\ &= \sum_{t=1}^{3} \frac{\sigma_t}{\left(1+R_t\right)^t}\end{aligned}$$

which is Eq. 21.20.

Appendix A
Statistical Tables

Table A.1 Probability function of the binomial distribution

The table shows the probability of x successes in n independent trials, each with probability of success p. For example, the probability of 4 successes in 8 independent trials, each with probability of success .35, is .1875

n	x	p									
		.05	*.10*	*.15*	*.20*	*.25*	*.30*	*.35*	*.40*	*.45*	*.50*
1	0	.9500	.9000	.8500	.8000	.7500	.7000	.6500	.6000	.5500	.5000
	1	.0500	.1000	.1500	.2000	.2500	.3000	.3500	.4000	.4500	.5000
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0312
	1	.2036	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1562
	2	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0000	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125

(continued)

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5,

Table A.1 (continued)

	4	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938
	6	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0000	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0037	.0078
8	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0312
	2	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4	.0004	.0046	.0815	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0312
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
11	0	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014	.0005
	1	.3293	.3835	.3248	.2362	.1549	.0932	.0518	.0266	.0125	.0054
	2	.0867	.2131	.2866	.2953	.2581	.1998	.1395	.0887	.0513	.0269
	3	.0137	.0710	.1517	.2215	.2581	.2568	.2254	.1774	.1259	.0806
	4	.0014	.0158	.0536	.1107	.1721	.2201	.2428	.2365	.2060	.1611
	5	.0001	.0025	.0132	.0388	.0803	.1321	.1830	.2207	.2360	.2256
	6	.0000	.0003	.0023	.0097	.0268	.0566	.0985	.1471	.1931	.2256

(continued)

Table A.1 (continued)

	7	.0000	.0000	.0003	.0017	.0064	.0173	.0379	.0701	.1128	.1611
	8	.0000	.0000	.0000	.0002	.0011	.0037	.0102	.0234	.0462	.0806
	9	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0052	.0126	.0269
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0021	.0054
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0005
12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002
	1	.3413	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029
	2	.0988	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161
	3	.0173	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537
	4	.0021	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208
	5	.0002	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934
	6	.0000	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256
	7	.0000	.0000	.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934
	8	.0000	.0000	.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208
	9	.0000	.0000	.0000	.0001	.0004	.0015	.0048	.0125	.0277	.0537
	10	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0068	.0161
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0029
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
13	0	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004	.0001
	1	.3512	.3672	.2774	.1787	.1029	.0540	.0259	.0113	.0045	.0016
	2	.1109	.2448	.2937	.2680	.2059	.1388	.0836	.0453	.0220	.0095
	3	.0214	.0997	.1900	.2457	.2517	.2181	.1651	.1107	.0660	.0349
	4	.0028	.0277	.0838	.1535	.2097	.2337	.2222	.1845	.1350	.0873
	5	.0003	.0055	.0266	.0691	.1258	.1803	.2154	.2214	.1989	.1571
	6	.0000	.0008	.0063	.0230	.0559	.1030	.1546	.1968	.2169	.2095
	7	.0000	.0001	.0011	.0058	.0186	.0442	.0833	.1312	.1775	.2095
	8	.0000	.0000	.0001	.0011	.0047	.0142	.0336	.0656	.1089	.1571
	9	.0000	.0000	.0000	.0001	.0009	.0034	.0101	.0243	.0495	.0873
	10	.0000	.0000	.0000	.0000	.0001	.0006	.0022	.0065	.0162	.0349
	11	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0012	.0036	.0095
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
14	0	.4877	.2288	.1028	.0440	.0178	.0068	.0024	.0008	.0002	.0001
	1	.3593	.3559	.2539	.1539	.0832	.0407	.0181	.0073	.0027	.0009
	2	.1229	.2570	.2912	.2501	.1802	.1134	.0634	.0317	.0141	.0056
	3	.0259	.1142	.2056	.2501	.2402	.1943	.1366	.0845	.0462	.0222
	4	.0037	.0348	.0998	.1720	.2202	.2290	.2022	.1549	.1040	.0611
	5	.0004	.0078	.0352	.0860	.1468	.1963	.2178	.2066	.1701	.1222
	6	.0000	.0013	.0093	.0322	.0734	.1262	.1759	.2066	.2088	.1833
	7	.0000	.0002	.0019	.0092	.0280	.0618	.1082	.1574	.1952	.2095
	8	.0000	.0000	.0003	.0020	.0082	.0232	.0510	.0918	.1398	.1833
	9	.0000	.0000	.0000	.0003	.0018	.0066	.0183	.0408	.0762	.1222
	10	.0000	.0000	.0000	.0000	.0003	.0014	.0049	.0136	.0312	.0611
	11	.0000	.0000	.0000	.0000	.0000	.0002	.0010	.0033	.0093	.0222
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0019	.0056
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0009
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

(continued)

Table A.1 (continued)

15	0	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000
	1	.3658	.3432	.2312	.1319	.0668	.0305	.0126	.0047	.0016	.0005
	2	.1348	.2669	.2856	.2309	.1559	.0916	.0476	.0219	.0090	.0032
	3	.0307	.1285	.2184	.2501	.2252	.1700	.1110	.0634	.0318	.0139
	4	.0049	.0428	.1156	.1876	.2252	.2186	.1792	.1268	.0780	.0417
	5	.0006	.0105	.0449	.1032	.1651	.2061	.2123	.1859	.1404	.0916
	6	.0000	.0019	.0132	.0430	.0917	.1472	.1906	.2066	.1914	.1527
	7	.0000	.0003	.0030	.0138	.0393	.0811	.1319	.1771	.2013	.1964
	8	.0000	.0000	.0005	.0035	.0131	.0348	.0710	.1181	.1647	.1964
	9	.0000	.0000	.0001	.0007	.0034	.0116	.0298	.0612	.1048	.1527
	10	.0000	.0000	.0000	.0001	.0007	.0030	.0096	.0245	.0515	.0916
	11	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0074	.0191	.0417
	12	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052	.0139
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0032
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
16	0	.4401	.1853	.0743	.0281	.0100	.0033	.0010	.0003	.0001	.0000
	1	.3706	.3294	.2097	.1126	.0535	.0228	.0087	.0030	.0009	.0002
	2	.1463	.2745	.2775	.2111	.1336	.0732	.0353	.0150	.0056	.0018
	3	.0359	.1423	.2285	.2463	.2079	.1465	.0888	.0468	.0215	.0085
	4	.0061	.0514	.1311	.2001	.2252	.2040	.1553	.1014	.0572	.0278
	5	.0008	.0137	.0555	.1201	.1802	.2099	.2008	.1623	.1123	.0667
	6	.0001	.0028	.0180	.0550	.1101	.1649	.1982	.1983	.1684	.1222
	7	.0000	.0004	.0045	.0197	.0524	.1010	.1524	.1889	.1969	.1746
	8	.0000	.0001	.0009	.0055	.0197	.0487	.0923	.1417	.1812	.1964
	9	.0000	.0000	.0001	.0012	.0058	.0185	.0442	.0840	.1318	.1746
	10	.0000	.0000	.0000	.0002	.0014	.0056	.0167	.0392	.0755	.1222
	11	.0000	.0000	.0000	.0000	.0002	.0013	.0049	.0142	.0337	.0667
	12	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0040	.0115	.0278
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0029	.0085
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
17	0	.4181	.1668	.0631	.0225	.0075	.0023	.0007	.0002	.0000	.0000
	1	.3741	.3150	.1893	.0957	.0426	.0169	.0060	.0019	.0005	.0001
	2	.1575	.2800	.2673	.1914	.1136	.0581	.0260	.0102	.0035	.0010
	3	.0415	.1556	.2359	.2393	.1893	.1245	.0701	.0341	.0144	.0052
	4	.0076	.0605	.1457	.2093	.2209	.1868	.1320	.0796	.0411	.0182
	5	.0010	.0175	.0668	.1361	.1914	.2081	.1849	.1379	.0875	.0472
	6	.0001	.0039	.0236	.0680	.1276	.1784	.1991	.1839	.1432	.0944
	7	.0000	.0007	.0065	.0267	.0668	.1201	.1685	.1927	.1841	.1484
	8	.0000	.0001	.0014	.0084	.0279	.0644	.1134	.1606	.1883	.1855
	9	.0000	.0000	.0003	.0021	.0093	.0276	.0611	.1070	.1540	.1855
	10	.0000	.0000	.0000	.0004	.0025	.0095	.0263	.0571	.1008	.1484
	11	.0000	.0000	.0000	.0001	.0005	.0026	.0090	.0242	.0525	.0944
	12	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0081	.0215	.0472
	13	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0021	.0068	.0182

(continued)

Table A.1 (continued)

	14	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
18	0	.3972	.1501	.0536	.0180	.0056	.0016	.0004	.0001	.0000	.0000
	1	.3763	.3002	.1704	.0811	.0338	.0126	.0042	0012	.0003	.0001
	2	.1683	.2835	.2556	.1723	.0958	.0458	.0190	.0069	.0022	.0006
	3	.0473	.1680	.2406	.2297	.1704	.1046	.0547	.0246	.0095	.0031
	4	.0093	.0700	.1592	.2153	.2130	.1681	.1104	.0614	.0291	.0117
	5	.0014	.0218	.0787	.1507	.1988	.2017	.1664	.1146	.0666	.0327
	6	.0002	.0052	.0301	.0816	.1436	.1873	.1941	.1655	.1181	.0708
	7	.0000	.0010	.0091	.0350	.0820	.1376	.1792	.1892	.1657	.1214
	8	.0000	.0002	.0022	.0120	.0376	.0811	.1327	.1734	.1864	.1669
	9	.0000	.0000	.0004	.0033	.0139	.0386	.0794	.1284	.1694	.1855
	10	.0000	.0000	.0001	.0008	.0042	.0149	.0385	.0771	.1248	.1669
	11	.0000	.0000	.0000	.0001	.0010	.0046	.0151	.0374	.0742	.1214
	12	.0000	.0000	.0000	.0000	.0002	.0012	.0047	.0145	.0354	.0708
	13	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0044	.0134	.0327
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0039	.0117
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0009	.0031
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0006
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
19	0	.3774	.1351	.0456	.0144	.0042	.0011	.0003	.0001	.0000	.0000
	1	.3774	.2852	.1529	.0685	.0268	.0093	.0029	.0008	.0002	.0000
	2	.1787	.2852	.2428	.1540	.0803	.0358	.0138	.0046	.0013	.0003
	3	.0533	.1796	.2428	.2182	.1517	.0869	.0422	.0175	.0062	.0018
	4	.0112	.0798	.1714	.2182	.2023	.1491	.0909	.0467	.0203	.0074
	5	.0018	.0266	.0907	.1636	.2023	.1916	.1468	.0933	.0497	.0222
	6	.0002	.0069	.0374	.0955	.1574	.1916	.1844	.1451	.0949	.0518
	7	.0000	.0014	.0122	.0443	.0974	.1525	.1844	.1797	.1443	.0961
	8	.0000	.0002	.0032	.0166	.0487	.0981	.1489	.1797	.1771	.1442
	9	.0000	.0000	.0007	.0051	.0198	.0514	.0980	.1464	.1771	.1762
	10	.0000	.0000	.0001	.0013	.0066	.0220	.0528	.0976	.1449	.1762
	11	.0000	.0000	.0000	.0003	.0018	.0077	.0233	.0532	.0970	.1442
	12	.0000	.0000	.0000	.0000	.0004	.0022	.0083	.0237	.0529	.0961
	13	.0000	.0000	.0000	.0000	.0001	.0005	.0024	.0085	.0233	.0518
	14	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0082	.0222
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0022	.0074
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
20	0	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000	.0000
	1	.3774	.2702	.1368	.0576	.0211	.0068	.0020	.0005	.0001	.0000
	2	.1887	.2852	.2293	.1369	.0669	.0278	.0100	.0031	.0008	.0002
	3	.0596	.1901	.2428	.2054	.1339	.0716	.0323	.0123	.0040	.0011

(continued)

Table A.1 (continued)

4	.0133	.0898	.1821	.2182	.1897	.1304	.0738	.0350	.0139	.0046
5	.0022	.0319	.1028	.1746	.2023	.1789	.1272	.0746	.0365	.0148
6	.0003	.0089	.0454	.1091	.1686	.1916	.1712	.1244	.0746	.0370
7	.0000	.0020	.0160	.0545	.1124	.1643	.1844	.1659	.1221	.0739
8	.0000	.0004	.0046	.0222	.0609	.1144	.1614	.1797	.1623	.1201
9	.0000	.0001	.0011	.0074	.0271	.0654	.1158	.1597	.1771	.1602
10	.0000	.0000	.0002	.0020	.0099	.0308	.0686	.1171	.1593	.1762
11	.0000	.0000	.0000	.0005	.0030	.0120	.0336	.0710	.1185	.1602
12	.0000	.0000	.0000	.0001	.0008	.0039	.0136	.0355	.0727	.1201
13	.0000	.0000	.0000	.0000	.0002	.0010	.0045	.0146	.0366	.0739
14	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0049	.0150	.0370
15	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0049	.0148
16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0046
17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011
18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Source: Reprinted from *Tables of the Binomial Probability Distribution* (1950), courtesy of the National Institute of Standards and Technology, Technology Administration, U.S. Department of Commerce

Table A.2 Poisson probabilities

For a given value of λ, entry indicates the probability of obtaining, a specified value of *X*

					λ					
X	*.1*	*.2*	*.3*	*.4*	*.5*	*.6*	*.7*	*.8*	*.9*	*1.0*
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
					λ					
X	*1.1*	*1.2*	*1.3*	*1.4*	*1.5*	*1.6*	*1.7*	*1.8*	*1.9*	*2.0*
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

(continued)

Table A.2 (continued)

					λ					
X	*2.1*	*2.2*	*2.3*	*2.4*	*2.5*	*2.6*	*2.7*	*2.8*	*2.9*	*3.0*
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1494
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680
5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027
10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008
11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
					λ					
X	*3.1*	*3.2*	*3.3*	*3.4*	*3.5*	*3.6*	*3.7*	*3.8*	*3.9*	*4.0*
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2180	.2158	.2125	.2087	.2046	.2001	.1954
4	.1734	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954
5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.2078	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
					λ					
X	*4.1*	*4.2*	*4.3*	*4.4*	*4.5*	*4.6*	*4.7*	*4.8*	*4.9*	*5.0*
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0280	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0014	.0016	.0019	.0022	.0026	.0030	.0034

(continued)

Table A.2 (continued)

13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002
					λ					
X	*5.1*	*5.2*	*5.3*	*5.4*	*5.5*	*5.6*	*5.7*	*5.8*	*5.9*	*6.0*
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0311	.0287	.0265	.0244	.0225	.0207	.0191	.0176	.0162	.0149
2	.0793	.0746	.0701	.0659	.0618	.0580	.0544	.0509	.0477	.0446
3	.1348	.1293	.1239	.1185	.1133	.1082	.1033	.0985	.0938	.0892
4	.1719	.1681	.1641	.1600	.1558	.1515	.1472	.1428	.1383	.1339
5	.1753	.1748	.1740	.1728	.1714	.1697	.1678	.1656	.1632	.1606
6	.1490	.1515	.1537	.1555	.1571	.1584	.1594	.1601	.1605	.1606
7	.1086	.1125	.1163	.1200	.1234	.1267	.1298	.1326	.1353	.1377
8	.0692	.0731	.0771	.0810	.0849	.0887	.0925	.0962	.0998	.1033
9	.0392	.0423	.0454	.0486	.0519	.0552	.0586	.0620	.0654	.0688
10	.0200	.0220	.0241	.0262	.0285	.0309	.0334	.0359	.0386	.0413
11	.0093	.0104	.0116	.0129	.0143	.0157	.0173	.0190	.0207	.0225
12	.0039	.0045	.0051	.0058	.0065	.0073	.0082	.0092	.0102	.0113
13	.0015	.0018	.0021	.0024	.0028	.0032	.0036	.0041	.0046	.0052
14	.0006	.0007	.0008	.0009	.0011	.0013	.0015	.0017	.0019	.0022
15	.0002	.0002	.0003	.0003	.0004	.0005	.0006	.0007	.0008	.0009
16	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003
17	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001
					λ					
X	*6.1*	*6.2*	*6.3*	*6.4*	*6.5*	*6.6*	*6.7*	*6.8*	*6.9*	*7.0*
0	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
1	.0137	.0126	.0116	.0106	.0098	.0090	.0082	.0076	.0070	.0064
2	.0417	.0390	.0364	.0340	.0318	.0296	.0276	.0258	0240	.0223
3	.0848	.0806	.0765	.0726	.0688	.0652	.0617	.0584	.0552	.0521
4	.1294	.1249	.1205	.1162	.1118	.1076	.1034	.0992	.0952	.0912
5	.1579	.1549	.1519	.1487	.1454	.1420	.1385	.1349	.1314	.1277
6	.1605	.1601	.1595	.1586	.1575	.1562	.1546	.1529	.1511	.1490
7	.1399	.1418	.1435	.1450	.1462	.1472	.1480	.1486	.1489	.1490
8	.1066	.1099	.1130	.1160	.1188	.1215	.1240	.1263	.1284	.1304
9	.0723	.0757	.0791	.0825	.0858	.0891	.0923	.0954	.0985	.1014
10	.0441	.0469	.0498	.0528	.0558	.0588	.0618	.0649	.0679	.0710
11	.0245	.0265	.0285	.0307	.0330	.0353	.0377	.0401	.0426	.0452
12	.0124	.0137	.0150	.0164	.0179	.0194	.0210	.0227	.0245	.0264
13	.0058	.0065	.0073	.0081	.0089	.0098	.0108	.0119	.0130	.0142
14	.0025	.0029	.0033	.0037	.0041	.0046	.0052	.0058	.0064	.0071
15	.0010	.0012	.0014	.0016	.0018	.0020	.0023	.0026	.0029	.0033
16	.0004	.0005	.0005	.0006	.0007	.0008	.0010	.0011	.0013	.0014
17	.0001	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0006
18	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002
19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

(continued)

Table A.2 (continued)

					λ					
X	*7.1*	*7.2*	*7.3*	*7.4*	*7.5*	*7.6*	*7.7*	*7.8*	*7.9*	*8.0*
0	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
1	.0059	.0054	.0049	.0045	.0041	.0038	.0035	.0032	.0029	.0027
2	.0208	.0194	.0180	.0167	.0156	.0145	.0134	.0125	.0116	.0107
3	.0492	.0464	.0438	.0413	.0389	.0366	.0345	.0324	.0305	.0286
4	.0874	.0836	.0799	.0764	.0729	.0696	.0663	.0632	.0602	.0573
5	.1241	.1204	.1167	.1130	.1094	.1057	.1021	.0986	.0951	.0916
6	.1468	.1445	.1420	.1394	.1367	.1339	.1311	.1282	.1252	.1221
7	.1489	.1486	.1481	.1474	.1465	.1454	.1442	.1428	.1413	.1396
8	.1321	.1337	.1351	.1363	.1373	.1382	.1388	.1392	.1395	.1396
9	.1042	.1070	.1096	.1121	.1144	.1167	.1187	.1207	.1224	.1241
10	.0740	.0770	.0800	.0829	.0858	.0887	.0914	.0941	.0967	.0993
11	.0478	.0504	.0531	.0558	.0585	.0613	.0640	.0667	.0695	.0722
12	.0283	.0303	.0323	.0344	.0366	.0388	.0411	.0434	.0457	.0481
13	.0154	.0168	.0181	.0196	.0211	.0227	.0243	.0260	.0278	.0296
14	.0078	.0086	.0095	.0104	.0113	.0123	.0134	.0145	.0157	.0169
15	.0037	.0041	.0046	.0051	.0057	.0062	.0069	.0075	.0083	.0090
16	.0016	.0019	.0021	.0024	.0026	.0030	.0033	.0037	.0041	.0045
17	.0007	.0008	.0009	.0010	.0012	.0013	.0015	.0017	.0019	.0021
18	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
19	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003	.0003	.0004
20	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001
					λ					
X	*8.1*	*8.2*	*8.3*	*8.4*	*8.5*	*8.6*	*8.7*	*8.8*	*8.9*	*9.0*
0	.0003	.0003	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001
1	.0025	.0023	.0021	.0019	.0017	.0016	.0014	.0013	.0012	.0011
2	.0100	.0092	.0086	.0079	.0074	.0068	.0063	.0058	.0054	.0050
3	.0269	.0252	.0237	.0222	.0208	.0195	.0183	.0171	.0160	.0150
4	.0544	.0517	.0491	.0466	.0443	.0420	.0398	.0377	.0357	.0337
5	.0882	.0849	.0816	.0784	.0752	.0722	.0692	.0663	.0635	.0607
6	.1191	.1160	.1128	.1097	.1066	.1034	.1003	.0972	.0941	.0911
7	.1378	.1358	.1338	.1317	.1294	.1271	.1247	.1222	.1197	.1171
8	.1395	.1392	.1388	.1382	.1375	.1366	.1356	.1344	.1332	.1318
9	.1256	.1269	.1280	.1290	.1299	.1306	.1311	.1315	.1317	.1318
10	.1017	.1040	.1063	.1084	.1104	.1123	.1140	.1157	.1172	.1186
11	.0749	.0776	.0802	.0828	.0853	.0878	.0902	.0925	.0948	.0970
12	.0505	.0530	.0555	.0579	.0604	.0629	.0654	.0679	.0703	.0728
13	.0315	.0334	.0354	.0374	.0395	.0416	.0438	.0459	.0481	.0504
14	.0182	.0196	.0210	.0225	.0240	.0256	.0272	.0289	.0306	.0324
15	.0098	.0107	.0116	.0126	.0136	.0147	.0158	.0169	.0182	.0194
16	.0050	.0055	.0060	.0066	.0072	.0079	.0086	.0093	.0101	.0109
17	.0024	.0026	.0029	.0033	.0036	.0040	.0044	.0048	.0053	.0058
18	.0011	.0012	.0014	.0015	.0017	.0019	.0021	.0024	.0026	.0029
19	.0005	.0005	.0006	.0007	.0008	.0009	.0010	.0011	.0012	.0014
20	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0005	.0006

(continued)

Table A.2 (continued)

21	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0002	.0003
22	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001
					λ					
X	*9.1*	*9.2*	*9.3*	*9.4*	*9.5*	*9.6*	*9.7*	*9.8*	*9.9*	*10*
0	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
1	.0010	.0009	.0009	.0008	.0007	.0007	.0006	.0005	.0005	.0005
2	.0046	.0043	.0040	.0037	.0034	.0031	.0029	.0027	.0025	.0023
3	.0140	.0131	.0123	.0115	.0107	.0100	.0093	.0087	.0081	.0076
4	.0319	.0302	.0285	.0269	.0254	.0240	.0226	.0213	.0201	.0189
5	.0581	.0555	.0530	.0506	.0483	.0460	.0439	.0418	.0398	.0378
6	.0881	.0851	.0822	.0793	.0764	.0736	.0709	.0682	.0656	.0631
7	.1145	.1118	.1091	.1064	.1037	.1010	.0982	.0955	.0928	.0901
8	.1302	.1286	.1269	.1251	.1232	.1212	.1191	.1170	.1148	.1126
9	.1317	.1315	.1311	.1306	.1300	.1293	.1284	.1274	.1263	.1251
10	.1198	.1210	.1219	.1228	.1235	.1241	.1245	.1249	.1250	.1251
11	.0991	.1012	.1031	.1049	.1067	.1083	.1098	.1112	.1125	.1137
12	.0752	.0776	.0799	.0822	.0844	.0866	.0888	.0908	.0928	.0948
13	.0526	.0549	.0572	.0594	.0617	.0640	.0662	.0685	.0707	.0729
14	.0342	.0361	.0380	.0399	.0419	.0439	.0459	.0479	.0500	.0521
15	.0208	.0221	.0235	.0250	.0265	.0281	.0297	.0313	.0330	.0347
16	.0118	.0127	.0137	.0147	.0157	.0168	.0180	.0192	.0204	.0217
17	.0063	.0069	.0075	.0081	.0088	.0095	.0103	.0111	.0119	.0128
18	.0032	.0035	.0039	.0042	.0046	.0051	.0055	.0060	.0065	.0071
19	.0015	.0017	.0019	.0021	.0023	.0026	.0028	.0031	.0034	.0037
20	.0007	.0008	.0009	.0010	.0011	.0012	.0014	.0015	.0017	.0019
21	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
22	.0001	.0001	.0002	.0002	.0002	.0002	.0003	.0003	.0004	.0004
23	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001
					λ					
X	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0010	.0004	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000
3	.0037	.0018	.0008	.0004	.0002	.0001	.0000	.0000	.0000	.0000
4	.0102	.0053	.0027	.0013	.0006	.0003	.0001	.0001	.0000	.0000
5	.0224	.0127	.0070	.0037	.0019	.0010	.0005	.0002	.0001	.0001
6	.0411	.0255	.0152	.0087	.0048	.0026	.0014	.0007	.0004	.0002
7	.0646	.0437	.0281	.0174	.0104	.0060	.0034	.0018	.0010	.0005
8	.0888	.0655	.0457	.0304	.0194	.0120	.0072	.0042	.0024	.0013
9	.1085	.0874	.0661	.0473	.0324	.0213	.0135	.0083	.0050	.0029
10	.1194	.1048	.0859	.0663	.0486	.0341	.0230	.0150	.0095	.0058
11	.1194	.1144	.1015	.0844	.0663	.0496	.0355	.0245	.0164	.0106
12	.1094	.1144	.1099	.0984	.0829	.0661	.0504	.0368	.0259	.0176
13	.0926	.1056	.1099	.1060	.0956	.0814	.0658	.0509	.0378	.0271
14	.0728	.0905	.1021	.1060	.1024	.0930	.0800	.0655	.0514	.0387
15	.0534	.0724	.0885	.0989	.1024	.0992	.0906	.0786	.0650	.0516

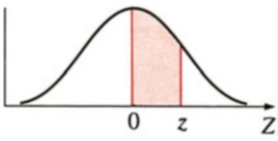
(continued)

Table A.2 (continued)

16	.0367	.0543	.0719	.0866	.0960	.0992	.0963	.0884	.0772	.0646
17	.0237	.0383	.0550	.0713	.0847	.0934	.0963	.0936	.0863	.0760
18	.0145	.0256	.0397	.0554	.0706	.0830	.0909	.0936	.0911	.0844
19	.0084	.0161	.0272	.0409	.0557	.0699	.0814	.0887	.0911	.0888
20	.0046	.0097	.0177	.0286	.0418	.0559	.0692	.0798	.0866	.0888
21	.0024	.0055	.0109	.0191	.0299	.0426	.0560	.0684	.0783	.0846
22	.0012	.0030	.0065	.0121	.0204	.0310	.0433	.0560	.0676	.0769
23	.0006	.0016	.0037	.0074	.0133	.0216	.0320	.0438	.0559	.0669
24	.0003	.0008	.0020	.0043	.0083	.0144	.0226	.0328	.0442	.0557
25	.0001	.0004	.0010	.0024	.0050	.0092	.0154	.0237	.0336	.0446
26	.0000	.0002	.0005	.0013	.0029	.0057	.0101	.0164	.0246	.0343
27	.0000	.0001	.0002	.0007	.0016	.0034	.0063	.0109	.0173	.0254
28	.0000	.0000	.0001	.0003	.0009	.0019	.0038	.0070	.0117	.0181
29	.0000	.0000	.0001	.0002	.0004	.0011	.0023	.0044	.0077	.0125
30	.0000	.0000	.0000	.0001	.0002	.0006	.0013	.0026	.0049	.0083
31	.0000	.0000	.0000	.0000	.0001	.0003	.0007	.0015	.0030	.0054
32	.0000	.0000	.0000	.0000	.0001	.0001	.0004	.0009	.0018	.0034
33	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0010	.0020
34	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0012
35	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0007
36	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0004
37	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
38	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
39	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

Source: Extracted from William H. Beyer, ed., *CRC Basic Statistical Tables* (Cleveland, Ohio: The Chemical Rubber Co., 1971)

Table A.3 The standardized normal distribution



The entries in this table are the probabilities that a standard normal random variable is between 0 and z (the shaded area)

	Second decimal place in z									
z	*.00*	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*.09*
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879

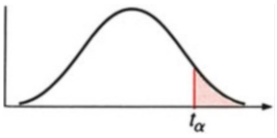
(continued)

Table A.3 (continued)

.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998									
4.0	.49997									
4.5	.499997									
5.0	.4999997									

Source: Reprinted from *Standard Mathematical Tables,* 15th ed., © CRC Press, Inc., Boca Raton, FL

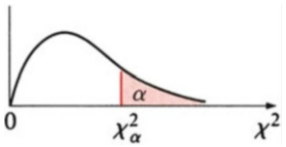
Table A.4 Critical values of t



Degrees of freedom ν	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.808
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Source: From M. Merrington, "Table of Percentage Points of the t-Distribution," *Biometrika*, 1941, 32,300. Reproduced by permission of the *Biometrika* trustees

Table A.5 Critical values of $\chi 2$



Degrees of freedom ν	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$
1	.0000393	.0001571	.0009821	.0039321	.0157908
2	.0100251	.0201007	.0506356	.102587	.210720
3	.0717212	.114832	.215795	.351846	.584375
4	.206990	.297110	.484419	.710721	1.063623
5	.411740	.554300	.831211	1.145476	1.61031
6	.675727	.872085	1.237347	1.63539	2.20413
7	.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518
11	2.60321	3.05347	3.81575	4.57481	5.57779
12	3.07382	3.57056	4.40379	5.22603	6.30380
13	3.56503	4.10691	5.00874	5.89186	7.04150
14	4.07468	4.66043	5.62872	6.57063	7.78953
15	4.60094	5.22935	6.26214	7.26094	8.54675
16	5.14224	5.81221	6.90766	7.96164	9.31223
17	5.69724	6.40776	7.56418	8.67176	10.0852
18	6.26481	7.01491	8.23075	9.39046	10.8649
19	6.84398	7.63273	8.90655	10.1170	11.6509
20	7.43386	8.26040	9.59083	10.8508	12.4426
21	8.03366	8.89720	10.28293	11.5913	13.2396
22	8.64272	9.54249	10.9823	12.3380	14.0415
23	9.26042	10.19567	11.6885	13.0905	14.8479
24	9.88623	10.8564	12.4011	13.8484	15.6587
25	10.5197	11.5240	13.1197	14.6114	16.4734
26	11.1603	12.1981	13.8439	15.3791	17.2919
27	11.8076	12.8786	14.5733	16.1513	18.1138
28	12.4613	13.5648	15.3079	16.9279	18.9392
29	13.1211	14.2565	16.0471	17.7083	19.7677
30	13.7867	14.9535	16.7908	18.4926	20.5992
40	20.7065	22.1643	24.4331	26.5093	29.0505
50	27.9907	29.7067	32.3574	34.7642	37.6886
60	35.5346	37.4848	40.4817	43.1879	46.4589
70	43.2752	45.4418	48.7576	51.7393	55.3290
80	51.1720	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7541	65.6466	69.1260	73.2912
100	67.3276	70.0648	74.2219	77.9295	82.3581

(continued)

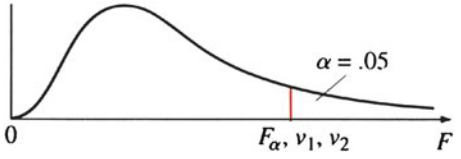
Table A.5 (continued)

Degrees of freedom ν	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381
4	7.77944	9.48773	11.1433	13.2767	14.8602
5	9.23635	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5346	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9871	18.3070	20.4831	23.2093	25.1882
11	17.2750	19.6751	21.9200	24.7250	26.7569
12	18.5494	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3621	24.7356	27.6883	29.8194
14	21.0642	23.6848	26.1190	29.1413	31.3193
15	22.3072	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8454	31.9999	34.2672
17	24.7690	27.5871	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5822
20	28.4120	31.4104	34.1696	37.5662	39.9968
21	29.6151	32.6705	35.4789	38.9321	41.4010
22	30.8133	33.9244	36.7807	40.2894	42.7956
23	32.0069	35.1725	38.0757	41.6384	44.1813
24	33.1963	36.4151	39.3641	42.9798	45.5585
25	34.3816	37.6525	40.6465	44.3141	46.9278
26	35.5631	38.8852	41.9232	45.6417	48.2899
27	36.7412	40.1133	43.1944	46.9630	49.6449
28	37.9159	41.3372	44.4607	48.2782	50.9933
29	39.0875	42.5569	45.7222	49.5879	52.3356
30	40.2560	43.7729	46.9792	50.8922	53.6720
40	51.8050	55.7585	59.3417	63.6907	66.7659
50	63.1671	67.5048	71.4202	76.1539	79.4900
60	74.3970	79.0819	83.2976	88.3794	91.9517
70	85.5271	90.5312	95.0231	100.425	104.215
80	96.5782	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.229
100	118.498	124.342	129.561	135.807	140.169

Source: From C. M. Thompson, "Tables of the Percentage Points of the χ2-Distribution," *Biometrika*, 1941, 32, 188–189. Reproduced by permission of the *Biometrika* trustees

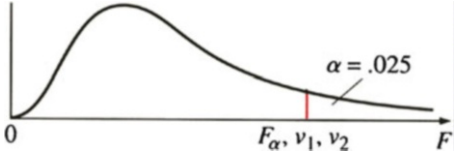
Table A.6 Critical values of F

For a particular combination of numerator and denominator degrees of freedom, entry represents the critical values of F corresponding to a specified upper tail area (α)



Denominator	*Numerator* ν_1																		
ν_2	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96

18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00



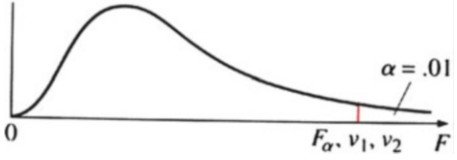
Denominator	*Numerator* ν_1																		
ν_2	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1,001	1,006	1,010	1,014	1,018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50

(continued)

Table A.6 (continued)

3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81

30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

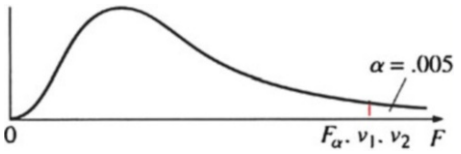


Denominator	*Numerator* ν_1																		
ν_2	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	∞
1	4,052	4999.5	5,403	5,625	5,764	5,859	5,928	5,982	6,022	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17

(continued)

Table A.6 (continued)

14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00



Deno-minator	*Numerator* ν_1																		
ν_2	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	∞
1	16,211	20,000	21,615	22,500	23,056	23,437	23,715	23,925	24,091	24,224	24,426	24,630	24,836	24,940	25,044	25,148	25,253	25,359	25,465
2	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5
3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83
4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.65	7.53	7.42	7.31	7.19	7.08
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34	4.23
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	.1.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69

(continued)

Table A.6 (continued)

21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	1.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69
120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43
∞	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00

Source: Reprinted from E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables far Statisticians*, 3rd ed., 1966. Reprinted by permission of the *Biometrika* trustees

Table A.7 Exponential function

c	e^{-c}	c	e^{-c}	c	e^{-c}
.00	1.000000	2.35	.095369	4.70	.009095
.05	.951229	2.40	.090718	4.75	.008652
.10	.904837	2.45	.086294	4.80	.008230
.15	.860708	2.50	.082085	4.85	.007828
.20	.818731	2.55	.078082	4.90	.007447
.25	.778801	2.60	.074274	4.95	.007083
.30	.740818	2.65	.070651	5.00	.006738
.35	.704688	2.70	.067206	5.05	.006409
.40	.670320	2.75	.063928	5.10	.006097
.45	.637628	2.80	.060810	5.15	.005799
.50	.606531	2.85	.057844	5.20	.005517
.55	.576950	2.90	.055023	5.25	.005248
.60	.548812	2.95	.052340	5.30	.004992
.65	.522046	3.00	.049787	5.35	.004748
.70	.496585	3.05	.047359	5.40	.004517
.75	.472367	3.10	.045049	5.45	.004296
.80	.449329	3.15	.042852	5.50	.004087
.85	.427415	3.20	.040762	5.55	.003887
.90	.406570	3.25	.038774	5.60	.003698
.95	.386741	3.30	.036883	5.65	.003518
1.00	.367879	3.35	.035084	5.70	.003346
1.05	.349938	3.40	.033373	5.75	.003183
1.10	.332871	3.45	.031746	5.80	.003028
1.15	.316637	3.50	.030197	5.85	.002880
1.20	.301194	3.55	.028725	5.90	.002739
1.25	.286505	3.60	.027324	5.95	.002606
1.30	.272532	3.65	.025991	6.00	.002479
1.35	.259240	3.70	.024724	6.05	.002358
1.40	.246597	3.75	.023518	6.10	.002243
1.45	.234570	3.80	.022371	6.15	.002133
1.50	.223130	3.85	.021280	6.20	.002029
1.55	.212248	3.90	.020242	6.25	.001930
1.60	.201897	3.95	.019255	6.30	.001836
1.65	.192050	4.00	.018316	6.35	.001747
1.70	.182684	4.05	.017422	6.40	.001661
1.75	.173774	4.10	.016573	6.45	.001581
1.80	.165299	4.15	.015764	6.50	.001503
1.85	.157237	4.20	.014996	6.55	.001430
1.90	.149569	4.25	.014264	6.60	.001360
1.95	.142274	4.30	.013569	6.65	.001294
2.00	.135335	4.35	.012907	6.70	.001231
2.05	.128735	4.40	.012277	6.75	.001171
2.10	.122456	4.45	.011679	6.80	.001114
2.15	.116484	4.50	.011109	6.85	.001059
2.20	.110803	4.55	.010567	6.90	.001008
2.25	.105399	4.60	.010052	6.95	.000959

(continued)

Table A.7 (continued)

c	e^{-c}	c	e^{-c}	c	e^{-c}
2.30	.100259	4.65	.009562	7.00	.000912
7.05	.000867	8.05	.000319	9.05	.000117
7.10	.000825	8.10	.000304	9.10	.000112
7.15	.000785	8.15	.000289	9.15	.000106
7.20	.000747	8.20	.000275	9.20	.000101
7.25	.000710	8.25	.000261	9.25	.000096
7.30	.000676	8.30	.000249	9.30	.000091
7.35	.000643	8.35	.000236	9.35	.000087
7.40	.000611	8.40	.000225	9.40	.000083
7.45	.000581	8.45	.000214	9.45	.000079
7.50	.000553	8.50	.000204	9.50	.000075
7.55	.000526	8.55	.000194	9.55	.000071
7.60	.000501	8.60	.000184	9.60	.000068
7.65	.000476	8.65	.000175	9.65	.000064
7.70	.000453	8.70	.000167	9.70	.000061
7.75	.000431	8.75	.000158	9.75	.000058
7.80	.000410	8.80	.000151	9.80	.000056
7.85	.000390	8.85	.000143	9.85	.000053
7.90	.000371	8.90	.000136	9.90	.000050
7.95	.000353	8.95	.000130	9.95	.000048
8.00	.000336	9.00	.000123	10.00	.000045

Table A.8 Random numbers

12651	61646	11769	75109	86996	97669	25757	32535	07122	76763
81769	74436	02630	72310	45049	18029	07469	42341	98173	79260
36737	98863	77240	76251	00654	64688	09343	70278	67331	98729
82861	54371	76610	94934	72748	44124	05610	53750	95938	01485
21325	15732	24127	37431	09723	63529	73977	95218	96074	42138
74146	47887	62463	23045	41490	07954	22597	60012	98866	90959
90759	64410	54179	66075	61051	75385	51378	08360	95946	95547
55683	98078	02238	91540	21219	17720	87817	41705	95785	12563
79686	17969	76061	83748	55920	83612	41540	86492	06447	60568
70333	00201	86201	69716	78185	62154	77930	67663	29529	75116
14042	53536	07779	04157	41172	36473	42123	43929	50533	33437
59911	08256	06596	48416	69770	68797	56080	14223	59199	30162
62368	62623	62742	14891	39247	52242	98832	69533	91174	57979
57529	97751	54976	48957	74599	08759	78494	52785	68526	64618
15469	90574	78033	66885	13936	42117	71831	22961	94225	31816
18625	23674	53850	32827	81647	80820	00420	63555	74489	80141
74626	68394	88562	70745	23701	45630	65891	58220	35442	60414
11119	16519	27384	90199	79210	76965	99546	30323	31664	22845
41101	17336	48951	53674	17880	45260	08575	49321	36191	17095

(continued)

Table A.8 (continued)

32123	91576	84221	78902	82010	30847	62329	63898	23268	74283
26091	68409	69704	82267	14751	13151	93115	01437	56945	89661
67680	79790	48462	59278	44185	29616	76531	19589	83139	28454
15184	19260	14073	07026	25264	08388	27182	22557	61501	67481
58010	45039	57181	10238	36874	28546	37444	80824	63981	39942
56425	53996	86245	32623	78858	08143	60377	42925	42815	11159
82630	84066	13592	60642	17904	99718	63432	88642	37858	25431
14927	40909	23900	48761	44860	92467	31742	87142	03607	32059
23740	22505	07489	85986	74420	21744	97711	36648	35620	97949
32990	97446	03711	63824	07953	85965	87089	11687	92414	67257
05310	24058	91946	78437	34365	82469	12430	84754	19354	72745
21839	39937	27534	88913	49055	19218	47712	67677	51889	70926
08833	42549	93981	94051	28382	83725	72643	64233	97252	17133
58336	11139	47479	00931	91560	95372	97642	33856	54825	55680
62032	91144	75478	47431	52726	30289	42411	91886	51818	78292
45171	30557	53116	04118	58301	24375	65609	85810	18620	49198
91611	62656	60128	35609	63698	78356	50682	22505	01692	36291
55472	63819	86314	49174	93582	73604	78614	78849	23096	72825
18573	09729	74091	53994	10970	86557	65661	41854	26037	53296
60866	02955	90288	82136	83644	94455	06560	78029	98768	71296
45043	55608	82767	60890	74646	79485	13619	98868	40857	19415
17831	09737	79473	75945	28394	79334	70577	38048	03607	06932
40137	03981	07585	18128	11178	32601	27994	05641	22600	86064
77776	31343	14576	97706	16039	47517	43300	59080	80392	63189
69605	44104	40103	95635	05635	81673	68657	09559	23510	95875
19916	52934	26499	09821	97331	80993	61299	36979	73599	35055
02606	58552	07678	56619	65325	30705	99582	53390	46357	13244
65183	73160	87131	35530	47946	09854	18080	02321	05809	04893
10740	98914	44916	11322	89717	88189	30143	52687	19420	60061
98642	89822	71691	51573	83666	61642	46683	33761	47542	23551
60139	25601	93663	25547	02654	94829	48672	28736	84994	13071

Source: From *A Million Random Digits with 100,000 Normal Deviates*, RAND (New York: The Fress Press) Copyright 1955 and 1983 by RAND. Used by permission

Table A.9 Cutoff points for the distribution of the Durbin-Watson test statistics

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
					$\alpha = .05$					
15	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21
16	1.10	1.37	.98	1.54	.86	1.73	.74	1.93	.62	2.15
17	1.13	1.38	1.02	1.54	.90	l.71	.78	1.90	.67	2.10
18	1.16	1.39	1.05	1.53	.93	1.69	.82	1.87	.71	2.06
19	1.18	1.40	1.08	1.53	.97	1.68	.86	1.85	.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	.93	1.81	.83	1.96

(continued)

Table A.9 (continued)

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
22	1.24	1.43	1.15	1.54	1.05	1.66	.96	1.80	.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	.99	1.79	.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78
	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
					$\alpha = .01$					
15	.81	1.07	.70	1.25	.59	1.46	.49	1.70	.39	1.96
16	.84	1.09	.74	1.25	.63	1.44	.53	1.66	.44	1.90
17	.87	1.10	.77	1.25	.67	1.43	.57	1.63	.48	1.85
18	.90	1.12	.80	1.26	.71	1.42	.61	1.60	.52	1.80
19	.93	1.13	.83	1.26	.74	1.41	.65	1.58	.56	1.77
20	.95	1.15	.86	1.27	.77	1.41	.68	1.57	.60	1.74
21	.97	1.16	.89	1.27	.80	1.41	.72	1.55	.63	1.71
22	1.00	1.17	.91	1.28	.83	1.40	.75	1.54	.66	1.69
23	1.02	1.19	.94	1.29	.86	1.40	.77	1.53	.70	1.67
24	1.04	1.20	.96	1.30	.88	1.41	.80	1.53	.72	1.66

(continued)

Table A.9 (continued)

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
25	1.05	1.21	.98	1.30	.90	1.41	.83	1.52	.75	1.65
26	1.07	1.22	1.00	1.31	.93	1.41	.85	1.52	.78	1.64
27	1.09	1.23	1.02	1.32	.95	.141	.88	1.51	.81	1.63
28	1.10	1.24	1.04	1.32	.97	1.41	.90	1.51	.83	1.62
29	1.12	1.25	1.05	1.33	.99	1.42	.92	1.51	.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	.94	1.51	.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	.96	1.51	.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	.98	1.51	.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

Source: From J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression, II," *Biometrika*, 1951, 30, 159–178. Reproduced by permission of the *Biometrika* trustees

Table A.10 Lower and upper critical values R for the runs test

n_2																			
n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
								Lower tail (α .025)											
2											2	2	2	2	2	2	2	2	2
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14
								Upper tail (α = .025)											
2																			
3																			
4				9	9														
5			9	10	10	11	11												
6			9	10	11	12	12	13	13	13	13								
7				11	12	13	13	14	14	14	14	15	15	15					
8				11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9					13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10					13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11					13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12					13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13						15	16	17	18	19	19	20	20	21	21	22	22	23	23
14						15	16	17	18	19	20	20	21	22	22	23	23	23	24
15						15	16	18	18	19	20	21	22	22	23	23	24	24	25
16							17	18	19	20	21	21	22	23	23	24	25	25	25
17							17	18	19	20	21	22	23	23	24	25	25	26	26
18							17	18	19	20	21	22	23	24	25	25	26	26	27
19							17	18	20	21	22	23	23	24	25	26	26	27	27
20							17	18	20	21	22	23	24	25	25	26	27	27	28

Source: Adapted from F. S. Swed and C. Eisenhart, *Ann. Math. Statist.*, 14, 1943, 83–86

Table A.11 Critical values of W in the Wilcoxon Matched-Pairs Signed-Rank test

For sample size n, the table shows, for selected probabilities, α, the numbers W_α, such that $P(W \leq W_\alpha) = \alpha$, where the distribution of the random variable W is that of the Wilcoxon test statistic under the null hypothesis

			α		
n	*.005*	*.010*	*.025*	*.050*	*.100*
4	0	0	0	0	1
5	0	0	0	1	3
6	0	0	1	3	4
7	0	1	3	4	6
8	1	2	4	6	9
9	2	4	6	9	11
10	4	6	9	11	15
11	6	8	11	14	18
12	8	10	14	18	22
13	10	13	18	22	27
14	13	16	22	26	32
15	16	20	26	31	37
16	20	24	30	36	43
17	24	28	35	42	49
18	28	33	41	48	56
19	33	38	47	54	63
20	38	44	53	61	70

Source: From R. L. McCornack, "Extended Tables of the Wilcoxon Matched Pairs Signed Rank Statistics," *Journal of the American Statistical Association, 60*(1965). Reprinted with permission from the Journal of the American Statistical Association. Copyright 1965 by the American Statistical Association. All rights reserved

Table A.12 Lower and upper critical values R_{n_1} and R_{n_2} of the Wilcoxon Rank-Sum test

	α		n_1						
n^2	One-tailed	Two-tailed	*4*	*5*	*6*	*7*	*8*	*9*	*10*
4	.05	.10	11.25						
	.025	.05	10.26						
	.01	.02	—.—						
	.005	.01	—.—						
5	.05	.10	12.28	19.36					
	.025	.05	11.29	17.38					
	.01	.02	10.30	16.39					
	.005	.01	—.—	15.40					
6	.05	.10	13.31	20.40	28.50				
	.025	.05	12.32	18.42	26.52				
	.01	.02	11.33	17.43	24.54				
	.005	.01	10.34	16.44	23.55				

(continued)

Table A.12 (continued)

	α		n_1						
n^2	One-tailed	Two-tailed	*4*	*5*	*6*	*7*	*8*	*9*	*10*
7	.05	.10	14.34	21.44	29.55	39.66			
	.025	.05	13.35	20.45	27.57	36.69			
	.01	.02	11.37	18.47	25.59	34.71			
	.005	.01	10.38	16.49	24.60	32.73			
8	.05	.10	15.37	23.47	31.59	41.71	51,85		
	.025	.05	14.38	21.49	29.61	38.74	49.87		
	.01	.02	12.40	19.51	27.63	35.77	45.91		
	.005	.01	11.41	17.53	25.65	34.78	43.93		
9	.05	.10	16.40	24.51	33.63	43.76	54.90	66,105	
	.025	.05	14.42	22.53	31.65	40.79	51.93	62,109	
	.01	.02	13.43	20.55	28.68	37.82	47.97	59,112	
	.005	.01	11.45	18.57	26.70	35.84	45.99	56,115	
10	.05	.10	17.43	26.54	35.67	45.81	56.96	69,111	82,128
	.025	.05	15.45	23.57	32.70	42.84	53.99	65,115	78,132
	.01	.02	13.47	21.59	29.73	39.87	49,103	61,119	74,136
	.005	.01	12.48	19.61	27.75	37.89	47,105	58,122	71,139

Source: Adapted from Table 1 of F. Wilcoxon and R. A. Wilcox, *Some Rapid Approximate Statistical Procedures*. Copyright © 1949. 1964 Lederle Laboratories, Division of American Cyanamid Company. All rights reserved. Reprinted with permission

Table A.13 Factors for control chart

	$\bar{X}$-charts					S-charts			R-charts					
n	*A*	A_2	A_3	c_4	B_3	B_4	B_5	B_6	d_2	d_3	D_1	D_2	D_3	D_4
2	2.121	1.880	2.659	.7979	0	3.267	0	2.606	1.128	.853	0	3.686	0	3.267
3	1.732	1.023	1.954	.8862	0	2.568	0	2.276	1.693	.888	0	4.358	0	2.574
4	1.500	.729	1.628	.9213	0	2.266	0	2.088	2.059	.880	0	4.698	0	2.282
5	1.342	.577	1.427	.9400	0	2.089	0	1.964	2.326	.864	0	4.918	0	2.114
6	1.225	.483	1.287	.9515	.030	1.970	.029	1.874	2.534	.848	0	5.078	0	2.004
7	1.134	.419	1.182	.9594	.118	1.882	.113	1.806	2.704	.833	.204	5.204	.076	1.924
8	1.061	.373	1.099	.9650	.185	1.815	.179	1.751	2.847	.820	.388	5.306	.136	1.864
9	1.000	.337	1.032	.9693	.239	1.761	.232	1.707	2.970	.808	.547	5.393	.184	1.816
10	.949	.308	.975	.9727	.284	1.716	.276	1.669	3.078	.797	.687	5.469	.223	1.777
11	.905	.285	.927	.9754	.321	1.679	.313	1.637	3.173	.787	.811	5.535	.256	1.744
12	.866	.266	.886	.9776	.354	1.646	.346	1.610	3.258	.778	.922	5.594	.283	1.717
13	.832	.249	.850	.9794	.382	1.618	.374	1.585	3.336	.770	1.025	5.647	.307	1.693
14	.802	.235	.817	.9810	.406	1.594	399	1.563	3.407	.763	1.118	5.696	.328	1.672
15	.775	.223	.789	.9823	.428	1.572	.421	1.544	3.472	.756	1.203	5.741	.347	1.653
16	.750	.212	.763	.9835	.448	1.552	.440	1.526	3.532	.750	1.282	5.782	.363	1.637
17	.728	.203	.739	.9845	.466	1.534	.458	1.511	3.588	.744	1.356	5.820	.378	1.622
18	.707	.194	.718	.9854	.482	1.518	.475	1.496	3.640	.739	1.424	5.856	.391	1.608
19	.688	.187	.698	.9862	.497	1.503	.490	1.483	3.689	.734	1.487	5.891	.403	1.597
20	.671	.180	.680	.9869	.510	1.490	.504	1.470	3.735	.729	1.549	5.921	.415	1.585
21	.655	.173	.663	.9876	.523	1.477	.516	1.459	3.778	.724	1.605	5.951	.425	1.575
22	.640	.167	.647	.9882	.534	1.466	.528	1.448	3.819	.720	1.659	5.979	.434	1.566
23	.626	.162	.633	.9887	.545	1.455	.539	1.438	3.858	.716	1.710	6.006	.443	1.557
24	.612	.157	.619	.9892	.555	1.445	.549	1.429	3.895	.712	1.759	6.031	.451	1.548
25	.600	.153	.606	.9896	.565	1.435	.559	1.420	3.931	.708	1.806	6.056	.459	4.541

Source: Copyright American Society for Testing and Materials. Reprinted with permission

Table A.14 Present value of $\$1$ $P = S_n(1 + r)^{-n}$

Years hence	1%	2%	4%	6%	8%	10%	12%	14%	15%	16%	18%	20%	22%	24%	25%	26%	28%	30%	35%	40%	45%	50%
1	.990	.980	.962	.943	.926	.909	.893	.877	.870	.862	.847	.833	.820	.806	.800	.794	.781	.769	.741	.714	.690	.667
2	.980	.961	.925	.890	.857	.826	.797	.769	.756	.743	.718	.694	.672	.650	.640	.630	.610	.592	.549	.510	.476	.444
3	.971	.942	.889	.840	.794	.751	.712	.675	.658	.641	.609	.579	.551	.524	.512	.500	.477	.455	.406	.364	.328	.296
4	.961	.924	.855	.792	.735	.683	.636	.592	.572	.552	.516	.482	.451	.423	.410	.397	.373	.350	.301	.260	.226	.198
5	.951	.906	.822	.747	.681	.621	.567	.519	.497	.476	.437	.402	.370	.341	.328	.315	.291	.269	.223	.186	.156	.132
6	.942	.888	.790	.705	.630	.564	.507	.456	.432	.410	.370	.335	.303	.275	.262	.250	.227	.207	.165	.133	.108	.088
7	.933	.871	.760	.665	.583	.513	.452	.400	.376	.354	.314	.279	.249	.222	.210	.198	.178	.159	.122	.095	.074	.059
8	.923	.853	.731	.627	.540	.467	.404	.351	.327	.305	.266	.233	.204	.179	.168	.157	.139	.123	.091	.068	.051	.039
9	.914	.837	.703	.592	.500	.424	.361	.308	.284	.263	.225	.194	.167	.144	.134	.125	.108	.094	.067	.048	.035	.026
10	.905	.820	.676	.558	.463	.386	.322	.270	.247	.227	.191	.162	.137	.116	.107	.099	.085	.073	.050	.035	.024	.017
11	.896	.804	.650	.527	.429	.350	.287	.237	.215	.195	.162	.135	.112	.094	.086	.079	.066	.056	.037	.025	.017	.012
12	.887	.788	.625	.497	.397	.319	.257	.208	.187	.168	.137	.112	.092	.076	.069	.062	.052	.043	.027	.018	.012	.008
13	.879	.773	.601	.469	.368	.290	.229	.182	.163	.145	.116	.093	.075	.061	.055	.050	.040	.033	.020	.013	.008	.005
14	.870	.758	.577	.442	.340	.263	.205	.160	.141	.125	.099	.078	.062	.049	.044	.039	.032	.025	.015	.009	.006	.003
15	.861	.743	.555	.417	.315	.239	.183	.140	.123	.108	.084	.065	.051	.040	.035	.031	.025	.020	.011	.006	.004	.002
16	.853	.728	.534	.394	.292	.218	.163	.123	.107	.093	.071	.054	.042	.032	.028	.025	.019	.015	.008	.005	.003	.002
17	.844	.714	.513	.371	.270	.198	.146	108	.093	.080	.060	.045	.034	.026	.023	.020	.015	.012	.006	.003	.002	.001
18	.836	.700	.494	.350	.250	.180	.130	.095	.081	.069	.051	.038	.028	.021	.018	.016	.012	.009	.005	.002	.001	.001
19	.828	.686	.475	.331	.232	.164	.116	.083	.070	.060	.043	.031	.023	.017	.014	.012	.009	.007	.003	.002	.001	
20	.820	.673	.456	.312	.215	.149	.104	.073	.061	.051	.037	.026	.019	.014	.012	.010	.007	.005	.002	.001	.001	
21	.811	.660	.439	.294	.199	.135	.093	.064	.053	.044	.031	.022	.015	.011	.009	.008	.006	.004	.002	.001		
22	.803	.647	.422	.278	.184	.123	.083	.056	.046	.038	.026	.018	.013	.009	.007	.006	.004	.003	.001	.001		
23	.795	.634	.406	.262	.170	.112	.074	.049	.040	.033	.022	.015	.010	.007	.006	.005	.003	.002	.001			
24	.788	.622	.390	.247	.158	.102	.066	.043	.035	.028	.019	.013	.008	.006	.005	.004	.003	.002	.001			
25	.780	.610	.375	.233	.146	.092	.059	.038	.030	.024	.016	.010	.007	.005	.004	.003	.002	.001	.001			
26	.772	.598	.361	.220	.135	.084	.053	.033	.026	.021	.014	.009	.006	.004	.003	.002	.002	.001				
27	.764	.586	.347	.207	.125	.076	.047	.029	.023	.018	.011	.007	.005	.003	.002	.002	.001	.001				
28	.757	.574	.333	.196	.116	.069	.042	.026	.020	.016	.010	.006	.004	.002	.002	.002	.001	.001				
29	.749	.563	.321	.185	.107	.063	.037	.022	.017	.014	.008	.005	.003	.002	.002	.001	.001	.001				
30	.742	.552	.308	.174	.099	.057	.033	.020	.015	.012	.007	.004	.003	.002	.001	.001	.001					

(continued)

Table A.14 (continued)

Years hence	1%	2%	4%	6%	8%	10%	12%	14%	15%	16%	18%	20%	22%	24%	25%	26%	28%	30%	35%	40%	45%	50%
40	.672	.453	.208	.097	.046	.022	.011	.005	.004	.003	.001	.001										
50	.608	.372	.141	.054	.021	.009	.003	.001	.001	.001												

Source: From Jerome Bracken and Charles J. Christenson, *Tables for Use in Analyzing Business Decisions.* 1965, reprinted by permission of Richard D. Irwin, Inc

Appendix B
Description of Data Sets

The following nine data sets used in the text are available on an IBM-compatible floppy disk, for instructors who request it. In addition, the data sets will be updated on disk each year.

Annual Macroeconomic Data (1960–2009)

The macroeconomic data included in this data set are GDP (gross domestic product), CPI (consumer price index), yield of 3-month T-bills, prime rate, private consumption, private investment, net exports, and government expenditures. The data set is also given in Table 2.2 in the text.

Financial Ratios for Two Pharmaceutical Companies (1990–2009)

The two pharmaceutical companies are Johnson & Johnson and Merck. The financial ratios included are the current ratio, inventory turnover, debt ratio (total debt/total assets), profit margin (net income/sales), return on assets (net income/total assets), P/E ratio, and payout ratio [dividend per share (DPS)/earnings per share (EPS)]. This data set is also given in Table 2.8 in the text. These data are also used in Appendix 3 of Chaps. 2 and 4 in the text.

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5,

EPS, DPS, PPS, and Rates of Return for Johnson & Johnson and Merck (1988–2009)

The data included in the first part of this data set are earnings per share (EPS), dividend per share (DPS), and price per share (PPS). At the far right of the data set is the S&P 500 Index. In the second part of the data set are the annual rates of return for JNJ, Merck, and the market. This information can also be found in Appendix 2 of Chap. 2 and Tables 2.3 and 2.4.

Annual JNJ Sales Data (1980–2010)

The set gives annual sales data for Johnson & Johnson from 1980 to 2010. The data are also presented in Table 18.7 in the text.

Quarterly EPS and Sales Data for Johnson & Johnson and IBM (2000–2010)

Included are quarterly EPS and sales data for Johnson & Johnson and IBM from the first quarter of 2000 to the fourth quarter of 2010. The EPS data for IBM can be found in Table 18.2. The EPS and sales data for Johnson & Johnson can also be found in Tables 18.5 and 18.10, respectively, in the text.

Monthly Rates of Return for Dow Jones 30 Companies (January 1990–December 2009)

This data set includes the monthly rates of return for the Dow Jones 30 and the S&P 500. The information is also given in Sect. 9.8 of the text.

The names of these 30 companies are:

3M Co.	Intel
Alcoa Inc.	IBM
American Express	Johnson & Johnson
AT&T	JP Morgan Chase & Co.
Bank of America	Kraft Foods
Boeing	McDonald's
Caterpillar Inc	Merck
Chevron	Microsoft
Cisco	Pfizer

(continued)

Coca-Cola	Procter & Gamble
E.I. du Pont de Nemours	Traveler's Companies Inc
Exxon	United Technologies Group
General Electric	Verizon
Hewlett-Packard	Walmart
Home Depot	Walt Disney

PPS, EPS, and DPS for Dow Jones 30 Companies (1990–2009)

This data set includes the annual PPS, EPS, and DPS data for Dow Jones 30 Companies. The annual S&P 500 Index is also included in this data set. The data can be found in Tables IV.1A–IV.1D of Project IV (Chap. 16).

Monthly Wilshire 5000 Equity Index (January 1989–January 1991)

The information given here is the value-weighted monthly Wilshire 5000 Equity Index. It can also be found in Table 19.10 in the text.

Monthly Rates of Return for the Value-Weighted Wilshire 5000 Equity Index (January 1989–January 1991)

Included are the percentage change in the price appreciation, dividend yield, and total rates of return of the value-weighted Wilshire 5000 Equity Index. The data are also given in Table 19.11 in the text.

Appendix C
Introduction to MINITAB 16

MINITAB 16 is a user-friendly statistics package. Students who are beginners in statistics will find that the application of MINITAB 16 in their statistics course will assist them in grasping statistics without greatly increasing their study load.

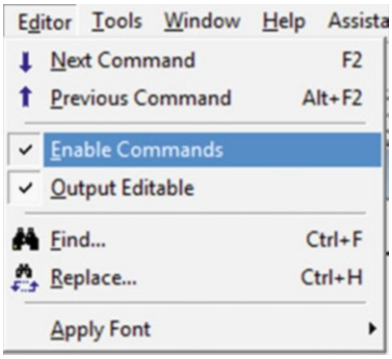
This appendix provides a brief description of the basic functions of MINITAB 16. With this basic knowledge, students can start to work with MINITAB 16.

General Description

MINITAB 16 is both a menu-driven and command-driven statistics package with more than 200 commands available. In this appendix, we will briefly look at the MINITAB 16 commands. Data are stored and processed in a worksheet, a table with rows and columns.

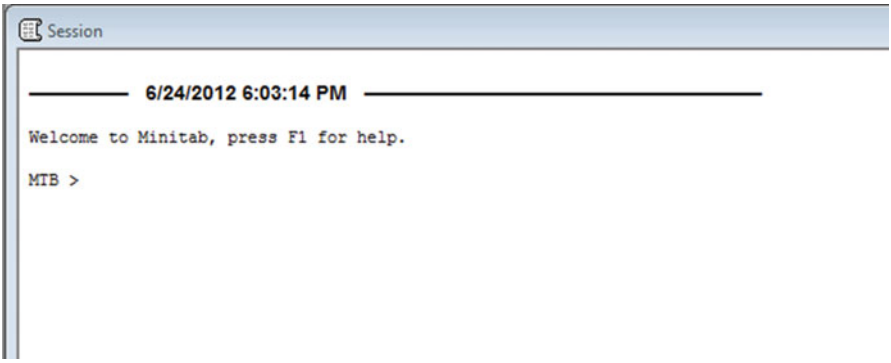
MINITAB 16 will accept words typed in upper- or lowercase letters or a combination of the two.

To enable MINITAB 16 commands, go the Editor menu and select the Enable Commands menu item.



C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5, © Springer Science+Business Media New York 2013

Doing this will show the MINITAB macro prompt in the Session windows. This is shown below.



Data Input

For MINITAB 16 to perform statistical computations, data must first be inputted. Data for each variable in your data set are stored in columns. There are two ways to enter data: READ and SET commands.

READ Commands

The form of the command is

```
READ C1
```

where C1 represents column 1. After typing this line, press ENTER. Then type the data, one number per line. The computer will prompt you after each entered line with

```
DATA>
```

When all the data have been input, type

```
END
```

For example, suppose you have the following four observations: 25, 33, 41, and 58. These data can be entered as follows:

```
MTB > read c1
DATA> 25
DATA> 31
DATA> 41
DATA> 58
DATA> end
```

If two or more groups of data (or variables) are to be read, you could type

```
READ C1 C2
```

or

```
READ C1 C2 C3
```

or

```
READ C1-C3
```

For example, suppose you have four observations for three variables:

```
variable 1: 13 17 21 11
variable 2: 23 27 32 20
variable 3: 35 31 42 47
```

The data can be entered as follows:

```
MTB > read c1-c3
DATA> 13 23 35
DATA> 17 27 31
DATA> 21 32 42
DATA> 11 20 37
DATA> end
```

SET Command

This command allows you to enter numbers consecutively on one or more lines for each variable. For instance, in the first example of the READ command, the data can be entered as follows:

```
MTB > set c1
DATA> 25 33 41 58
DATA> end
```

The second example of READ command can be entered as follows:

```
MTB > set c1
DATA> 13 17 21 11
DATA> end
MTB > set c2
DATA> 23 27 32 20
DATA> end
MTB > set c3
DATA> 35 31 42 37
DATA> end
```

Data Corrections

Suppose that after entering the data you find an error. The following three instructions allow you to correct errors:

1. The LET command enables you to replace an erroneous enter. For example,

   ```
   LET C2(3) = 7
   ```

2. The DELETE command simply erases the data you specify. For example,

   ```
   DELETE 3:6 C3
   ```

3. The INSERT command lets you add new material. For example,

   ```
   INSERT BETWEEN ROWS 4 AND 5 OF COLUMN C2 AND C3
   ```

```
DATA> 13 17
DATA>END
```

Insert a new row of data between rows 4 and 5. The new data are 13 and 17.

Output

To check to see if entered data have been inputted correctly on the screen, type

```
PRINT C1
```

or

```
PRINT C1 C2 C3
```

or

```
PRINT C1-C3
```

To print out the results of your statistical operations on paper, type PAPER before you enter the print commands. To stop printing, type NOPAPER after you get your printout.

Savings Data

To save a data set, type SAVE 'a:filename'. Once the data set has been saved, you can retrieve it at any time with the following command:

```
RETRIEVE 'a:filename'
```

Other Commands

To create new variables from existing ones, use the LET command. For example,

```
LET C4=C2+C3
```

It creates a variable that is the sum of the values stored in the second and third columns, and that variable is stored in column 4.

The MINITAB symbols for common arithmetic operations are

```
+ add
-subtract
* multiply
/ divide
** exponent (raise to a power)
```

```
  To erase entire columns, type
  ERASE C1 C2 C3
```

or

```
  ERASE C1-C3
```

When you have completed your work in MINITAB, type

```
  STOP
```

Appendix D
Introduction to SAS: Microcomputer Version

SAS is a powerful statistics package for manipulating data and performing varied statistical analyses. It is designed for larger data bases.

This appendix will provide a brief description of the basic functions of SAS. With this basic knowledge, students can easily understand SAS programs written by others and start to write SAS programs themselves.

General Description

SAS is normally run in batch mode, rather than interactively as MINITAB is. Therefore, we first put all the command statements into a file (or program) and then submit the entire file to SAS for processing. SAS will perform the requested analyses and return two files: a log of the program (the SAS log), with notes and error messages, and a list of the results from the analyses. The SAS log is a record of everything that you do in your SAS program. Original program statements are identified by line numbers. Interspersed with SAS statements are messages from SAS. These messages might begin with the words NOTE, INFO, WARNING, ERROR, or an error number, and they might refer to a SAS statement by its line number in the log. Note that *every* statement in the SAS program must end with a semicolon. SAS will accept words typed in upper- or lowercase letters or a combination of the two.

Data Input

For SAS to perform statistical computations, data must first be inputted. The first command in any SAS program is the DATA command. For example, the command

```
DATA SALES;
```

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5,

will create a data set called "SALES." To read in the data, we then use the INPUT command, which tells SAS how the data values are arranged on the data lines and what the variable names are. An example of an INPUT command is

```
INPUT SALES REGION $:
```

This command informs SAS that you are going to read in two variables, called SALES and REGION. The listing of variable names in the INPUT statement tells SAS that the data are arranged on the data lines in the order listed, with at least one space between values. The dollar sign after REGION tells SAS that the variable region contains alphabetic characters.

After the INPUT command comes the CARDS statement, which tells SAS where to start reading, followed by the data to be read. For example, we may have sales and cost data for several regions listed as follows:

Region	Cost	Sales
East	4,325	5,647
West	5,941	7,103
South	2,387	3,492
North	3,762	4,481

Our entire data input will be as follows:

```
DATA SALES;
INPUT SALES COST REGION $;
CARDS;
5647 4325 EAST
7103 5941 WEST
3492 2387 SOUTH
4481 3762 NORTH
```

Data Modifying

To create new variables from existing ones, simply specify the appropriate formula, using the following symbols for the standard arithmetic operations:

```
+ add
– subtract
*multiply
/ divide
** raise to a power
```

Here are some examples:

```
PROFIT = SALES – COST;
MARGIN = (SALES – COST)/COST;
```

Analyzing the Data

SAS procedures (nicknamed PROCs) are used to process data in SAS data sets. There are procedures for all kinds of analyses from printing the input data to simple statistics to more complicated statistical analyses. The SAS procedures are written after the data lines. The following will introduce three SAS procedures: PROC PRINT, PROC ANOVA, and PROC REG.

PROC PRINT

The PROC PRINT statement asks SAS to print out the data values in the data set just created. The word PROC signals the beginning of a PROC step, a series of statements that describes the analysis to be performed. The word print names the SAS procedure we want to use.

PROC ANOVA

This procedure statement must be followed by a statement identifying the treatment variable and the model. Two examples are follows:

Completely Randomized Design

```
DATA EXAMPLE1;
INPUT X T;
CARDS;
25 1
27 1
31 1
32 2
35 2
30 2
27 3
32 3
48 3
18 4
23 4
29 4
;
PROC ANOVA;
CLASS T;
MODEL X = T;
RUN;
```

Randomized Block Design

```
DATA EXAMPLE2;
INPUT X T B;
CARDS;
25 1 1
27 1 2
31 1 3
32 2 1
35 2 2
30 2 3
27 3 1
32 3 2
48 3 3
18 4 1
23 4 2
29 4 3
;
PROC ANOVA;
CLASS T B;
MODEL X = T B;
RUN;
```

PROC REG

The REG procedure is used to perform both simple and multiple regressions. The procedure statement is followed by the MODEL statement, which specifies the dependent and independent variables. For example, the commands and data input for the multiple regression for Table 14.13 are as follows:

```
DATA TAB1413;
INPUT Y X1 X2 X3;
CARDS;
260.3 5 3 4
286.1 7 5 2
279.4 6 3 3
410.8 9 4 4
438.2 12 6 1
315.3 8 3 4
565.1 11 7 3
570.0 16 8 2
426.1 13 4 3
315.0 7 3 4
403.6 10 6 1
220.5 4 4 1
```

```
343.6 9 4 3
644.6 17 8 4
520.4 19 7 2
329.5 9 3 2
426.0 11 6 4
343.2 8 3 3
450.4 13 5 4
421.8 14 5 2
245.6 7 4 4
503.3 16 6 3
375.7 9 5 3
265.5 5 3 3
620.6 18 6 4
450.5 18 5 3
270.1 5 3 2
368.0 7 6 2
556.1 12 7 1
570.0 13 6 4
318.5 8 4 3
260.2 6 3 2
667.0 16 8 2
618.3 19 8 2
525.3 17 7 4
332.2 10 4 3
393.2 12 5 3
283.5 8 3 3
376.2 10 5 4
481.8 12 5 2
;
PROC REG;
MODEL Y = X1 X2 X3/DW;
RUN;
```

The regression results are as follows:

The SAS System 06:20 Monday, July 12, 2012
The REG Procedure
Model: MODEL1
Dependent Variable: Y
Analysis of Variance

Source	DF	Sum of squares	Mean square	F value	Pr > F
Model	3	527209	175736	89.05	<.0001
Error	36	71044	1973.44286		
Corrected Total	39	598253			

Root MSE	44.42345	R-Square	0.8812
Dependent Mean	411.28750	Adj R-Sq	0.8714
Coeff Var	10.80107		

Parameter Estimates

Variable	DF	Parameter estimate	Standard error	t value	Pr > |t|
Intercept	1	31.15039	34.17505	0.91	0.3681
X1	1	12.96816	2.73723	4.74	<.0001
*X*2	1	41.24562	7.28011	5.67	<.0001
X3	1	11.52425	7.69118	1.50	0.1428
		Durbin-Watson D	2.104		
		Number of Observations	40		
		First Order Autocorrelation	−0.083		

Appendix E
Useful Formulas in Statistics

Chapter 4

A. Measures of central tendency:

1. Sample arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.1}$$

2. Sample geometric mean:

$$\bar{x}_g = \left(x_1 \times x_2 \times x_3 \times \ldots \times x_n\right)^{1/n} \tag{4.3}$$

3. Grouped mean:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i m_i}{\sum_{i=1}^{k} f_i} \tag{4.17}$$

4. Median:

$$m = L + \frac{(N/2 - F)}{f}(U - L) \tag{4.18}$$

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5,

B. Measures of dispersion:

1. Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \tag{4.5}$$

2. Sample variance:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{4.7}$$

3. Sample mean absolute deviation:

$$\text{MAD} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \quad \text{or} \quad \frac{\sum_{i=1}^{n} (x_i - \text{Md}_s)}{n} \tag{4.9}$$

4. Coefficient of variation:

$$\text{CV}_x = \frac{S}{\bar{x}} \tag{4.12}$$

5. Population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}} \tag{4.6}$$

6. Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{4.8}$$

7. Population variance for frequency distribution:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \mu)^2}{N} \tag{4.19}$$

8. Sample variance for frequency distribution:

$$S^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \bar{x})^2}{n-1} \tag{4.21}$$

C. Population measures of skewness:

1. Skewness:

$$\mu_3 = \frac{\sum_{i=1}^{N} (x_i - \mu)^3}{N} \tag{4.15}$$

2. Population coefficient of skewness:

$$\text{CS} = \frac{\mu_3}{\sigma_3} \tag{4.16}$$

Chapter 5

1. Probability of Event A:

$$P_r(A) = \frac{\text{Number favourable outcomes}}{\text{Total number of outcomes}} \tag{5.1}$$

2. Union of Events A and B:

$$P_r(A \cup B) = P_r(A) + P_r(B) - P_r(A \cap B) \tag{5.5}$$

3. Intersection of Events A and B:

$$P_r(A \cap B) = P_r(A) + P_r(B) - P_r(A \cup B) \tag{5.6}$$

4. Probability of Complements:

$$P_r(E \cup \bar{E}) = P_r(E) + P_r(\bar{E}) = 1 \tag{5.9}$$

5. Conditional Probability:

$$P_r(B|A) = \frac{P_r(B \cap A)}{P_r(A)} \tag{5.13}$$

6. Bayes' Theorem:

$$P_r(B|A) = \frac{P_r(A \cap B)}{P_r(A)} = \frac{P_r(A|B)P_r(B)}{P_r(A)} \tag{5.19}$$

7. Number of Permutations of n-things taken r at a time

$$_nP_r = \frac{n!}{(n-r)!} \tag{5A.4}$$

8. Number of Combinations r objects can be selected from n

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{5.12}$$

Chapter 6

A. Binomial distribution:

$$P_r(x\,\text{sucess}\,|n\,\text{trials}) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n \tag{6.9}$$

$$\mu_x = np \tag{6.11}$$

$$\sigma_x^2 = np(1-p) \tag{6.12}$$

B. Hypergeometric distribution:

$$P_r(x\,\text{success}\,|\,n\,\text{trials}) = \frac{C_x^h C_{n-x}^{N-h}}{C_n^N} \tag{6.13}$$

$$\mu_x = n\left[\frac{h}{N}\right] \tag{6.14}$$

$$\sigma_x^2 = \left(\frac{N-n}{N-1}\right)\left[n\left(\frac{h}{N}\right)\left(1-\frac{h}{N}\right)\right] \tag{6.15}$$

C. Poisson distribution:

$$P_r(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}\,\text{for}\,x = 0, 1, 2, 3\,\text{and}\,\lambda > 0 \tag{6.16}$$

$$\mu_x = \lambda \tag{6.17a}$$

$$\sigma_x^2 = \lambda \tag{6.17b}$$

D. Expected value for a discrete random variable:

$$\mu = E(X) = \sum_{i=1}^{N} x_i P(x_i) \tag{6.3}$$

E. Variance for a discrete random variable:

$$\sigma^2 = E[x_i - \mu]^2 = \sum_{i=1}^{N} (x_i - \mu)^2 P(x_i) \tag{6.4}$$

F. Covariance for two discrete random variables:

$$\mathrm{Cov}(X, Y) \equiv \sigma_{X,Y} = \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) P_i(X_i, Y_i) \tag{6.25}$$

G. Correlation coefficient for two discrete random variables:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{6.27}$$

H. Marginal probability:

1.

$$P(x_i) = \sum_{j=1}^{m} P(x_i, y_j) \qquad i = 1, \ldots, n \tag{6.20}$$

2.

$$P(y_j) = \sum_{i=1}^{n} P(x_i, y_j) \qquad i = 1, \ldots, n \tag{6.21}$$

I. Conditional probability:

$$P(x|y) = \frac{P(x, y)}{P(y)} \tag{6.22}$$

Chapter 7

1. Standard normal variable:

$$z = \frac{x - \mu_x}{\sigma_x} \tag{7.4}$$

2. Normal approximation of binomial:

$$z = \frac{x - np}{\sqrt{np(1-p)}} \tag{7.10}$$

3. Normal approximation of Poisson:

$$z = \frac{x - \lambda}{\sqrt{\lambda}} \tag{7.14}$$

4. Probability that x lies between a and b:

   (a) Discrete random variable

$$P(a \leq x \leq b) = \sum_{i=1}^{b} P(x_i) - \sum_{i=1}^{a} P(x_i) \tag{7.1}$$

   (b) Continuous variable

$$P(a) = \int_{a}^{b} f(x)dx \tag{7A.2}$$

5. Mean of lognormal distribution:

$$\mu_X = e^{\mu + \sigma^2/2} \tag{7.6}$$

6. Variance of lognormal distribution:

$$\sigma_X^2 = e^{2\mu + \sigma^2(e^{\sigma^2} - 1)} \tag{7.7}$$

Chapter 8

1. Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{8.1}$$

2. Sample variance:

$$s_x^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2 \qquad \text{(footnote 2)}$$

3. Standard deviation of the sample mean:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \tag{8.4}$$

4. Mean for a sample proportion:

$$n_{\hat{p}} = \frac{np}{n} = p \tag{8.9}$$

5. Standard deviation of a sample proportion:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{8.10}$$

Chapter 9

1. Uniform probability density function:

$$f(x) = \begin{cases} 1/(b-a) & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases} \tag{9.1}$$

2. Uniform cumulative distribution function:

$$P(X \le x) = F(x) = \begin{cases} 0 & \text{if } x \le a \\ (x-a)/(b-a) & \text{if } a \le x \le b \\ 1 & \text{if } x > b \end{cases} \tag{9.2}$$

3. Mean and variance for uniform distribution:

$$\mu_x = \frac{a+b}{2}, \quad \sigma_x^2 = \frac{b-a}{\sqrt{12}} \tag{9.3}$$

4. Exponential probability density function:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0, \\ & \lambda > 0 \\ 0 & \text{if } t < 0 \end{cases} \tag{9.16}$$

5. Exponential cumulative distribution function:

$$\begin{aligned} F(t) &= 1 - e^{-\lambda t}, \quad t \geq 0 \\ &= 0 \qquad\quad t < 0 \end{aligned} \tag{9.17}$$

6. Mean and variance for exponential distribution:

$$E(T) = \frac{1}{\lambda}, \qquad \text{var}(T) = \frac{1}{\lambda^2} \tag{9.18}$$

7. t statistic:

$$t = \frac{\bar{x} - \mu}{s_x/\sqrt{n}} \tag{9.4}$$

8. Chi-square distribution:

$$\frac{(n-1)S_x^2}{\sigma_x^2} = \frac{\sum (x_i - \bar{x})^2}{\sigma_x^2} \sim \chi_n^2 \tag{9.8}$$

9. F distribution:

$$\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} = F \tag{9.14}$$

Chapter 10

1. $$\text{Unbiasedness}: \ E(\hat{\theta}) = \theta \tag{10.1}$$

2. $$\text{Relative Efficiency} = \frac{V(\theta_2)}{V(\theta_1)} \tag{10.2}$$

3. Confidence intervals for the mean:

   (a) Variance known:

$$1-\alpha = P_r\left[\bar{X} - Z_{\alpha/2}\left(\frac{\sigma_X}{\sqrt{n}}\right) < \mu < \bar{X} + Z_{\alpha/2}\left(\frac{\sigma_X}{\sqrt{n}}\right)\right] \tag{10.7}$$

   (b) Variance unknown:

$$1-\alpha = P_r\left[\bar{X} - t_{n-1,\alpha/2}\left(\frac{S_X}{\sqrt{n}}\right) < \mu < \bar{X} + t_{n-1,\alpha/2}\left(\frac{S_X}{\sqrt{n}}\right)\right] \tag{10.9}$$

4. Confidence interval for a proportion:

$$1-\alpha = P_r\left[\hat{P}_x - Z_{\alpha/2}\sqrt{\frac{\hat{P}_x(1-\hat{P}_x)}{n}} < \hat{P}_x + Z_{\alpha/2}\sqrt{\frac{\hat{P}_x(1-\hat{P}_x)}{n}}\right] \tag{10.11}$$

5. Confidence interval for the variance:

$$1-\alpha = P_r\left[\frac{(n-1)s_X^2}{\chi^2_{v,\alpha/2}} < \sigma_X^2 < \frac{(n-1)s_X^2}{\chi^2_{v,1-\alpha/2}}\right] \tag{10.13}$$

Chapter 11

1. Testing one mean-known variance:

$$\frac{\bar{x}-\mu_0}{\frac{\sigma_x}{\sqrt{n}}} = z \tag{11.3}$$

2. Testing one mean-unknown variance with large sample:

$$\frac{\bar{x}-\mu_0}{\frac{s_x}{\sqrt{n}}} = z$$

3. Testing one mean-unknown variance with small sample:

$$\frac{\bar{x}-\mu_0}{\frac{s_x}{\sqrt{n}}} = t_{n-1} \tag{10.8}$$

4. Testing a proportion:

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}=z \tag{10.10}$$

5. Testing the difference between two means — small sample:

$$\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}=t \tag{11.12}$$

$$\text{df}=n_1+n_2-2 \tag{11.13}$$

6. Testing the difference between two means — large sample:

$$\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_2}}}=z \tag{11.5}$$

Chapter 12

A. Testing the equality of three or more means:

1. $$F=\frac{\text{SST}/(m-1)}{\text{SSW}/(n-m)}=\frac{\text{MST}}{\text{MSW}} \tag{12.7}$$

2. $$\text{SST}=\sum_{j}^{m} n_j(\bar{x}_j-\bar{x})^2 \tag{12.4}$$

3. $$\text{SSW}=\sum_{j=1}^{\text{m}}\sum_{i=1}^{nj}(x_{ij}-\bar{x}_j)^2 \tag{12.5}$$

B. Two-way ANOVA:

1. Treatment effect:

$$\frac{\text{SST}/(J-1)}{\text{SSE}/(I-1)(J-1)}\sim F_{(J-1),(I-1)(J-1)}$$

$$\text{SSB} = \sum_{i=1}^{I} JK(\bar{x}_{i..} - \bar{x})^2 \tag{12.11}$$

$$\text{SST} = \sum_{j=1}^{J} IK(\bar{x}_{.j.} - \bar{x})^2 \tag{12.12}$$

$$\text{SSE} = \text{TSS-SST-SSB} \tag{12.13}$$

2. Block effect:

$$\frac{\text{SSB}/(I-1)}{\text{SSE}/(I-1)(J-1)} \sim F_{(I-1),(I-1)(J-1)}$$

3. Interaction effect:

$$\frac{\text{SSI}/(J-1)(I-1)}{\text{SSE}/IJ(K-1)}$$

$$\text{SSI} = K \sum_i \sum_j (\bar{x}_{ij.} - \bar{x}_{.j.} - \bar{x}_{i..} + \bar{x})^2$$

$$\text{SSE} = \sum_i \sum_j \sum_k (\bar{x}_{ijk} - \bar{x}_{ij.})^2 \tag{12.15}$$

4. Goodness of fit test:

$$x_{k-1}^2 = \sum_{i=1}^{k} \frac{(f_i^0 - f_i^e)^2}{f_i^e} \tag{12.17}$$

Chapter 13

1. Single regression equation:

$$y_i = a + bx_i + e_i \tag{13.6}$$

2. Least squares estimates:

(a) $$a = \bar{y} - b\bar{x} \tag{13.11}$$

(b)
$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{13.13}$$

.

3. Total variation = unexplained variation + explained variation (13.17)

4. Coefficient of determination:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{13.21}$$

5. Adjusted coefficient of determination:

$$\bar{R}^2 = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} \tag{13.22}$$

6. Standard error of the residual:

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} \tag{13.20}$$

7. The population correlation coefficient:

$$\rho = \frac{\text{Cov}(x,\ y)}{\sigma_x \sigma_y} \tag{13.23}$$

8. The sample correlation coefficient:

$$r = \frac{S_{xy}}{S_x S_y} \tag{13.24}$$

Chapter 14

1. Significance test for $b = 0$:

$$t_{n-2} = \frac{b - b_0}{S_b} \tag{14.4}$$

$$S_b^2 = \frac{S_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{14.2}$$

2. Significance test for $a \neq 0$:

$$t_{n-2} = \frac{a-0}{S_a} \tag{14.5}$$

$$S_a^2 = S_e^2 \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{14.1}$$

3. F-test for the significance of b:

$$F_{1,n-2} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{\text{MSR}}{\text{MSE}} \tag{14.9}$$

4. Significance test for r:

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{14.10}$$

5. Conditional expectation interval:

$$\hat{y}_{n+1} \pm t_{\alpha/2.n-2} S_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{14.21}$$

6. Prediction interval:

$$\hat{y}_{n+1} \pm t_{\alpha/2.n-2} S_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{14.22}$$

Chapter 15

1. Multiple regression model:

$$y_i = a + bx_{1i} + b_2 x_{2i} + e_i \tag{15.2}$$

2. Regression coefficients with two independent variables:

$$b_1 = \frac{\left(\sum_{i=1}^{n} x_{i_2}^2 - n\bar{x}_2^2\right)\left[\sum_{i=1}^{n} x_{i1}y_i - n\bar{x}_1\bar{y}\right] - \left(\left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\bar{x}_1\bar{x}_2\right)\left(\sum_{i=1}^{n} x_{i2}y_i - n\bar{x}_2\bar{y}\right)\right)}{\left(\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2\right) - \left(\sum_{i=1}^{n} x_{i2} - n\bar{x}_2^2\right) - \left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\bar{x}_1\bar{x}_2\right)^2} \tag{15.7}$$

$$b_2 = \frac{\left(\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2\right)\left(\sum_{i=1}^{n} x_{i2}y_i - n\bar{x}_2\bar{y}\right) - \left(\left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\bar{x}_1\bar{x}_2\right)\left(\sum_{i=1}^{n} x_{i1}y_i - n\bar{x}_1\bar{y}\right)\right)}{\left(\sum_{i=1}^{n} x_{i1}^2 - n\bar{x}_1^2\right) - \left(\sum_{i=1}^{n} x_{i2} - n\bar{x}_2^2\right) - \left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\bar{x}_1\bar{x}_2\right)^2} \tag{15.8}$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \tag{15.10}$$

3. Coefficient of determination:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{15.15}$$

4. Adjusted coefficient of determination:

$$\bar{R}^2 = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} = 1 - (1 - R^2)\frac{(n-1)}{n-k-1} \tag{15.16}$$

F-ratio:

$$F_{k,n-k-t} = \left(\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}\right)\left(\frac{n-k-1}{k}\right) \tag{15.19}$$

Chapter 16

1. Variance inflationary factor (VIF):

$$\text{VIF} = \frac{1}{1 - R_i^2} \tag{16.6}$$

2. Durbin-Watson statistics:

$$\mathrm{DW} = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{16.11}$$

3. Durbin's H:

$$\mathrm{H} = \left[1 - \frac{d}{2}\right] \sqrt{\frac{n}{1 - nV(\gamma)}} \tag{16.22}$$

4. Quadratic regression model:

$$y_i = \alpha + b_1 X_1 + b_2 X_1^2 + e_i \tag{16.16}$$

5. Lagged dependent variable model:

$$y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + \gamma y_{t-1} + e_i \tag{16.20}$$

6. Dummy variable model:

$$y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \gamma D_{1i} + e_i \tag{16.25}$$

7. Interaction variable model:

$$y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 (X_{1t} \times X_{2t}) + e_i \tag{16.28}$$

Chapter 17

1. Mann-Whitney U test:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 - 1)}{2} - R_1 \tag{17.5}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_1 \tag{17.6}$$

2. Kruskal-Wallis test:

$$K \left[\frac{12}{n(n+1)}\right] \sum_{i=1}^{c} \left[\frac{R_i^2}{n_i}\right] - 3(n+1) \tag{17.9}$$

3. Spearman's rank correlation:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{17.10}$$

4. t-statistic for rank correlation:

$$t = \frac{r_s}{\sqrt{(1-r_s^2)/(n-2)}} \tag{17.11}$$

5. Runs test:

$$Z_t = \frac{R - \mu_R}{\sigma_R} \tag{17.12}$$

where

$$\mu_R = \frac{2n_1n_2}{n_1n_2} + 1$$

$$\sigma_R^2 = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}$$

Chapter 18

1. Additive model for time-series component:

$$X_t = T_t + C_t + S_t + I_t \tag{18.1}$$

2. Multiplicative model for time-series component:

$$X_t = T_t S_t C_t I_t \tag{18.2}$$

3. k-term moving average:

$$Z_t = \frac{1}{k}\sum_{i=0}^{k-1} x_{t-i}, \qquad t = (k, \ldots, n) \tag{18.4}$$

4. Exponential smoothing:

$$S_{t+1} = \alpha X_t + (1-\alpha)S_t \tag{18.15}$$

5. Linear time trend:

$$X_t = \alpha + \beta t + \varepsilon_t \tag{18.11}$$

6. Mean square error (MSE):

$$\text{MSE} = \frac{\sum_{t=1}^{n} (x_t - \hat{x}_t)^2}{n}$$

7. The Holt-Winters Forecasting Model:

$$s_t = \alpha x_t + (1 - \alpha)(s_{t-1} + T_{t-1}) \tag{18.19a}$$

$$T_t = \beta(s_t - s_{t-1}) + (1 - \beta)T_{t-1} \tag{18.19b}$$

8. p-th order autoregressive process:

$$\hat{x}_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots + a_p x_{t-p} \tag{18.23}$$

Chapter 19

A. Price index	Formula	Comments
1. Simple aggregate price index	$I_t = \frac{\sum_{i=1}^{n} P_{ti}}{\sum_{i=1}^{n} P_{0i}} \times 100$ (19.1)	Does not consider the relative importance of each component Unit can affect the index value
2. Simple relative price index	$I_t = \frac{\sum_{i=1}^{n} (P_{ti}/P_{t0})}{n} \times 100$ (19.2)	All commodities are treated equally The index does not reflect the importance of individual commodities
3. Laspeyres index	$I_t = \frac{\sum_{i=1}^{n} P_{ti} Q_{0i}}{\sum_{i=1}^{n} P_{0i} Q_{0i}} \times 100$ (19.6)	Tends to give more weight to those items that show a dramatic price increase

(continued)

A.	Price index	Formula		Comments
4.	Paasche index	$I_t = \frac{\sum_{i=1}^{n} P_{ti}Q_{ti}}{\sum_{i=1}^{n} P_{0i}Q_{ti}} \times 100$	(19.7)	Provides a more up-to-date estimate of total expenses than the Laspeyres Complicated to update because it uses reference year quantities Tends to understate a price increase and overstate a price decrease
5.	Fisher's ideal price index	$FI_t = \sqrt{\frac{\sum P_{ti}Q_{ti}}{\sum P_{0i}Q_{ti}} \frac{\sum P_{ti}Q_{0i}}{\sum P_{0i}Q_{0i}}} \times 100$	(19.8)	Compromise between Laspeyres and Paasche

B.	Quantity indexes	Formula		Comments
1.	Laspeyres (LQ)	$I_t = \frac{\sum_{i=1}^{n} Q_{ti}P_{0i}}{\sum_{i=1}^{n} Q_{0i}P_{0i}} \times 100$	(19.9)	Gives more weight to those commodities that show a dramatic quantity increase
2.	Paasche (PQ)	$I_t = \frac{\sum_{i=1}^{n} Q_{ti}P_{ti}}{\sum_{i=1}^{n} Q_{0i}P_{ti}} \times 100$	(19.10)	More up to date than Laspeyres
3.	Fisher's quantity index	$FI_t = \sqrt{(LQ)(PQ)}$	(19.11)	Compromise between Laspeyres and Paasche
4.	Value indexes	$I_t = \frac{\sum_{i=1}^{n} Q_{ti}P_{ti}}{\sum_{i=1}^{n} Q_{0i}P_{0i}} \times 100$	(19.12)	

Chapter 20

A. Simple random sampling:

1. Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{20.1}$$

2. Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{20.2}$$

3. Estimated variance for population mean:

$$\sigma_{\bar{x}}^2 = \frac{s^2}{n} \times \frac{N-n}{N} \tag{20.3}$$

4. Confidence interval:

$$\bar{X} - z_{\alpha/2}\hat{\sigma}_{\bar{x}} < \mu < \bar{X} + z_{\alpha/2}\hat{\sigma}_{\overline{X}} \tag{20.4}$$

B. Simple random sampling for proportions:

1. Estimated variance for population mean:

$$\hat{\sigma}_p^2 = \frac{\hat{p}(1-\hat{p})}{n-1} \times \frac{N-n}{N} \tag{20.5}$$

2. Confidence interval:

$$\hat{p} - z_{\alpha/2}\hat{\sigma}_{\hat{p}} < p < \hat{p} + z_{\alpha/2}\hat{\sigma}_{\hat{p}} \tag{20.6}$$

C. Stratified random sampling:

1. Unbiased estimate for population mean:

$$\bar{X}_{st} = \sum_{j=1}^{H} W_j \bar{X}_j \tag{20.7}$$

2. Estimated variance for population mean:

$$\hat{\sigma}_{\bar{x}}^2 = \sum_{j=1}^{n} W_j^2 \hat{\sigma}_{\bar{X}_j}^2 \tag{20.8}$$

D. Sample size:

1. Sample size for simple random sampling:

$$n = \frac{N\sigma^2}{(N-1)\sigma_{\bar{X}}^2 + \sigma^2} \tag{20.9}$$

2. Sample size for stratified random sampling:

$$n = \frac{\sum_{j=1}^{H} N_j S_j^2}{N\sigma_{\bar{x}st}^2 + \frac{1}{N}\sum_{j=1}^{H} N_j S_j^2} \tag{20.13}$$

3. Optimal proportion for jth strata:

$$n_j = \frac{N_j \sigma_j}{\sum_{j=1}^{K} N_j S_j} \cdot n \tag{20.14}$$

4. Optimal allocation for total sample:

$$n = \frac{\frac{1}{N}\left(\sum_{j=1}^{K} N_j \sigma_j\right)^2}{N\sigma_{\bar{x}}^2 + \frac{1}{N}\sum_{j=1}^{K} N_j S_j^2} \tag{20.15}$$

E. Cluster sampling:

1. Estimated population mean:

$$\hat{\mu} = \frac{\sum_{i=1}^{m} N_i \bar{x}_i}{(\bar{N})(m)} \tag{20.17}$$

2. Estimated variance for:

$$\sigma_{\hat{\mu}}^2 = \left[\frac{M-m}{M\,m\,\bar{N}^2}\right] \times \frac{\sum_{i=1}^{m} N_i^2 (X_i - \bar{X})^2}{m-1} \tag{20.19}$$

3. Confidence interval:

$$\bar{X} - z_{\alpha/2}\sigma_{\hat{\mu}}^2 < \mu < \bar{X} + z_{\alpha/2}\sigma_{\hat{\mu}}^2 \tag{20.20}$$

F. Ratio method:

$$\hat{x}_r = \frac{\bar{x}}{\bar{y}} Y \tag{20.21}$$

G. Regression method:

$$\hat{\mu}_x = \bar{x} + b\left(\mu_y - \bar{y}\right) \tag{20.22}$$

Chapter 21

1. Expected monetary value:

$$\text{EMV}(A_i) = \sum_{j=1}^{M} P_j M_{ij} \qquad (i = 1, 2, \ldots, H) \tag{12.1}$$

2. Expected utility:

$$\text{E}[U(A_i)] = \sum_{j=1}^{m} P_j U_{ij} \qquad (i = 1, 2, \ldots, n) \tag{12.2}$$

3. Generalized Bayes model:

$$P(S_i|I) = \frac{P(I|S_i)P(S_i)}{\sum P(I|S_i)P(S_i)} \tag{21.3}$$

Appendix F
Important Finance and Accounting Topics

A. *Financial Ratio Analysis*

1. Static analysis: Appendix 3 of Chap. 2, Example 3.5, Appendix 3 of Chap. 4, and Application 12.3
2. Dynamic analysis: Appendix 1 of Chap. 16

B. *Interest Rate, Inflation Rate, and Term Structure of Interest Rate*

1. Interest rate: Example 2.2, Example 3.3, Example 4.17, Application 4.3, Application 5.1, and Application 7.4
2. Inflation rate: Example 4.10 and Application 19.3
3. Term structure of interest rate: Appendix 2 of Chap. 16

C. *Stock Rates of Return and Portfolio Analysis*

1. Stock rates of return: Application 2.1, Appendix 2 of Chap. 2, Example 3.5, Example 4.2, Example 4.3, Example 4.4, Example 4.11, Example 4.14, Example 4.15, Application 4.2, Example 6.22, Application 7.1. Section 9.8, Application 12.4, and Application 17.1
2. Portfolio analysis: Example 6.21 and Appendix 1 of Chap. 13

D. *Market Model and Capital Asset Pricing Model*

1. Market model: Application 14.2 and Example 21.11
2. Capital asset pricing model: Sect. 21.7 and Appendix 2 of Chap. 21

E. *Utility Analysis*: Sect 21.4

F. *Capital Budgeting Decision*: Application 7.3, Example 21.9, Sect. 21.8, and Appendix 4 of Chap. 21

G. *Option Pricing Models*: Appendix 2 of Chap. 6, Appendices 2 and 3 of Chap. 7, Appendix 5 of Chap. 9, Appendix 4 of Chap. 13, and Appendix 1 of Chap. 19

H. *Stock Market Indexes and Hedge Ratios*: Example 1.14, Sect. 19.6, and Appendix 2 of Chap. 19

I. *Dividend Behavior Model*: Example 16.7

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5,

J. *Cash Management Model*: Appendix 1 of Chap. 10
K. *Cost–Volume–Profit Analysis*: Application 7.2
L. *Investment Performance Measure*: Sect. 21.7

Index

C.-F. Lee et al., *Statistics for Business and Financial Economics*,
DOI 10.1007/978-1-4614-5897-5, © Springer Science+Business Media New York 2013